The performance of medical record review as an instrument for measuring and improving patient safety

The performance of medical record review as an instrument for measuring and improving patient safety

Dorthe Odyl Klein

The studies presented in this dissertation were conducted at Maastricht University Medical Centre+, at the department of Clinical epidemiology and Health Technology Assessment (KEMTA).

Printing of this thesis was financially supported by Medirede (trade name of Mediround)

Dorthe Odyl Klein, Margraten 2019

Cover design: Renneke Korte-Doornekamp www.delinden.info
Layout: Danny de Haan
Printed by: Gildeprint
ISBN: 9789463238373

The performance of medical record review as an instrument for measuring and improving patient safety

PROEFSCHRIFT

Ter verkrijging van de graad van doctor aan de Universiteit Maastricht,
op gezag van de Rector Magnificus, Prof. Dr. Rianne M. Letschert,
volgens besluit van het College van Decanen,
in het openbaar te verdedigen
op woensdag 6 november 2019 om 14.00 uur

door

Dorthe Odyl Klein

Promotoren

Prof. Dr. R.P. Koopmans
Prof. Dr. M.H. Prins

Copromotor

Dr. R.J.M.W. Rennenberg

Beoordelingscommissie

Prof. Dr. C.A.B. Webers (voorzitter)
Prof. Dr. C.G. Faber
Prof. Dr. S.E. Geerlings (Amsterdam UMC)
Dr. L.J.G.G. Panis
Prof. Dr. C. Wagner (Amsterdam UMC)

Table of contents

Abbreviations

AE	Adverse Event
COOP	Committee investigating deceased patients (Commissie Onderzoek Overleden Patiënten)
CNN	Convolutional Neural Networks
DALY	Disability-adjusted life years
FT	Fast Text neural network
HMPS	Harvard Medical Practice Study
GTT	Global Trigger Tool
IHI	Institute for Healthcare Improvement
HSMR	Hospital Standardized Mortality Rate
IOM	Institute of Medicine
NIVEL	Dutch institute for health services research
MRR	Medical Record Review
MUMC+	Maastricht University Medical Centre+
NB	Naïve Bayes
NLP	Natural Language Processing
NPV	Negative predictive value
PABAK	Prevalence And Bias Adjusted Kappa
PPV	Positive Predictive Value
SVM	Support Vector Machine
WHO	World Health Organization

Chapter 1

Introduction

General introduction

Imagine being a patient yourself in need of hospital services. You would expect a hospital to be safe and its service of high quality. According to the World Health Organization (WHO), patient safety can be defined as the prevention of errors and adverse effects to patients associated with health care.[1] However, reports in the past decades showed that around 1 in 10 admitted patients experience care related harm.[2-6] Converted to global numbers these are approximately 42.7 million adverse events (AEs). An AE could lead to a temporary or permanent injury, but also to the death of a patient.[7-9] Together, these AEs result in 23 million disability lost life years (DALYs) per year worldwide.[10]

Although the numbers mentioned seem large, we should realise that only a part of these AEs are preventable.[8,11-13] Moreover, it is impossible to investigate all cases which means that the numbers mentioned above are extrapolations made from a sample of cases.

However, since these reports with large estimated numbers of AEs worldwide were published, many countries decided to implement measures to improve patient safety. They used various instruments to detect and improve quality and safety and created specific safety programs.[11,14-18] In short, it was put on the social agenda. Quality and safety of care are thus important for health care institutions, policy makers and healthcare professionals but most importantly patients.

In order to be able to improve quality and safety, organizations have to measure their basic performance, identify bottlenecks and find ways to implement improvement strategies. The result of these actions should then preferably lead to measurable improvement of the parameters used to measure quality and safety. However, this is actually a surrogate marker for improvement because in the end the patient should profit. In this thesis we zoom in on the measurement of the basic performance and identification of bottlenecks as an important start of the improvement process.

To get an indication of quality, several hospital parameters are nowadays registered. For example, hospitals in the Netherlands are participating in, on average, 45 quality registries, nineteen quality accreditations and seven patient experience reports.[19-24] To get this registration done, in the Netherlands alone approximately 80 million euros are spend on a yearly base.[25]

Some examples of these quality indicators and safety parameters are waiting time, number of patients with decubitus, number of fall incidents, excess length of stay, AEs, complications, hospital standardized mortality rate (HSMR), number of reoperations etcetera.[26-31]
To measure these indicators and parameters correctly, health care institutions have several instruments at their disposal. One of them, which will be discussed here in more detail, is AE analysis. This is a commonly used instrument in patient safety research. AEs can be detected in many different ways, for example with medical record review, voluntary data reporting, incident reporting, patient interviews, complaints registries and *morbidity* and *mortality* conferences.[32,33-40]

Medical record review is one of the most frequently used methods for identifying AEs.[5,13,41-43] This is also in our hospital the key method for evaluating potentially preventable AEs and a basic measurement of performance.

Approach to quality control in Maastricht UMC+ using medical record review

In 2008, the project ''preventing medical injury'' was started in Maastricht UMC+. The founders of this project had the impression that the HSMR was sensitive to improper and unintended influences and the nature of the hospital population. The HSMR would therefore not give useful insights into the causes of AEs and possible points for improvement.[44,45] Moreover; it only has a signal function and does not yield information about where to improve the quality of care and what to change.[46,47]

Hence, the aim of the project was to decrease and prevent unintended medical harm to patients in the MUMC+. They chose to use systematic retrospective evaluation of the records of all deceased patients.
To learn from preventable medical harm and to prevent it in the future, the goal was also to stimulate an open culture in which caregivers are open to communicate about, and learn from (the consequences of) their own actions. Because, in their opinion, the most severe outcome of AEs was death, they chose to use the records of deceased patients.

The board of directors of the Maastricht UMC+ installed a committee with the assignment to evaluate all medical records of patients who died during their stay in the hospital. This committee was named: Committee Investigating Deceased Patients (In Dutch: Commissie Onderzoek Overleden Patiënten; COOP).

Medical record review method

Reviewing all medical records is extremely time consuming. Therefore, usually samples are taken of prespecified patient groups but also trigger tools are used.[17,48,49] A trigger is a warning signal that an AE could be present in the record thus limiting the number of records to be investigated and still yield a useful number of AEs to change policy or learn about an institutions weakness concerning quality and safety.

The most well-known trigger tools are the Global trigger tool (GTT) and the Harvard Medical Practice Study (HMPS).[5,50-52] The HMPS is more often deployed in research settings, whereas the GTT is usually deployed for reviews in patient safety work.[17,53,54]

The records of the deceased patients are first explored by a team of trained nurses for the presence of triggers (see table 1). Subsequently, a committee consisting of medical specialists from all major disciplines analysed the records to search for AEs. Both the screeners and the specialists were not time restricted. All results were saved using software which was specially designed for this purpose.
The bottlenecks identified are then used to advise on improvement measures for the management of the departments. Between 2008 and 2015 we noticed regularly that in our feedback to the departments the quality of our assessment method was an often-discussed issue. Because this hampered the implementation of improvement measures we decided to measure the quality of the instrument we used.

Therefore, this PhD project was started to assess the quality and performance of our medical record review method looking for potential preventable AEs.

An AE was defined as an unintended outcome arising from the (non)-action of a caregiver and/or the health care system with damage to the patient resulting in temporary or permanent disability or death of the patient.[55] If a potentially preventable AE was suspected, this was discussed with the involved medical department. Finally, the committee decided on the definite presence of an AE, its potential preventability and the contribution to death. For the purpose of the study in this thesis, we used the committee result as a gold standard for AEs. Of all the available trigger systems we chose to use the HMPS list because this was already used by the Netherlands Institute for health services research (NIVEL) in their evaluation of hospital performance.[13,56,57] Our screeners were already trained because they participated in the NIVEL screenings in using this list which made implementation easy. We hypothesized that this list would also be redundant in deceased patients.[58] Due to the very nature of deceased subjects and experience with the triggers the list was slightly adapted. Trigger 1 (patient was admitted before (in the previous 12 months) for a reason related to the current admission) was adapted to a shorter period (in the previous 3 months) because analysis of previous years showed this trigger was not discriminative for potentially preventable AEs. The 12-month cut-off contained a large number of patients with planned chemotherapy or planned second stage operations. Furthermore, two other triggers were not applicable in a deceased population and were omitted (the trigger regarding unplanned transfer to another acute care hospital and inappropriate discharge to home).

Table 1: A list of the triggers which are used in the MUMC+

Description of the triggers
Unplanned readmission after discharge from index admission within 3 months
Hospital-incurred patient injury (temporarily or lasting)
Adverse drug reaction
Unplanned transfer to ICU
Unplanned return to the operating room
Unplanned removal, injury or repair of an organ during surgery
Healthcare related infection or sepsis
Other complications such as CVA/pulmonary embolism
Development of neurological deficit not present on admission
(initial) unexpected and/or sudden death (not palliative care)
Cardiac or respiratory arrest
Injury related to abortion or delivery
Dissatisfaction with care
Documentation indicating litigation
Other patient complications

Although this method has similarities with the original method (which has been studied by other groups), there are also some differences.[5,11,59-62] First, we focus only on deceased patients. Second, we don't use reviewer pairs since these don't improve the reliability.[61,63] Instead, one reviewer evaluates the medical record and then presents his findings to the other members of the committee. During this meeting it is decided whether or not the event was an AE. Third, we use internal reviewers for the fact that we believe they have the best knowledge of the medical process in our centre.[64] Finally; the final decision on the presence of an AE is made after the committee weighed the opinion of the involved care givers.

Thesis outline

The focus of this thesis was to assess the quality of medical record review as an instrument to increase patient safety in clinical care eventually. An overview of these sub studies is shown in figure 1.

Part I: What is already known?

Trigger tools are frequently used as a screening instrument to select cases for extensive review. Because searching for triggers, but even more scrutinizing medical records, is a time consuming and therefore costly procedure we assumed that these methods were thoroughly studied for their performance. Therefore, in the first part of the thesis we focus on what is already known about medical record review, more specific concerning the trigger tools which are often used combined with medical record review.

In **chapter two** we provided a systematic narrative review on the two most frequently used trigger tools, the HMPS and the GTT. We evaluated their performance regarding 7 criteria, which were created by the WHO to evaluate screening methods on their performance.[65] These criteria consist of an evaluation of the effectiveness in capturing the extent of harm, availability of reliable data and an evaluation of the costs of AEs. Furthermore, its effectiveness in influencing policy, hospital and local safety procedures and outcomes was investigated. Finally, the synergy with other domains of quality of care was assessed.

Part II: Assessing the quality of medical record review

After the evaluation of the current knowledge of medical record review in part I, we decided to perform further evaluations of the quality of MRR. The method used in our hospital is slightly different, since it evaluates only AEs in deceased patients. Also, the final conclusion on the presence of an AE is decided after a group discussion.

In this part we evaluate the quality of medical record review. Therefore, we have split the evaluation in the trigger part (performed by the nurses) and the assessment of the AEs (performed by the committee).

In **chapter three**, we performed an analysis on the predictive value of the triggers.

In **chapter four**, we evaluated the internal reproducibility of the triggers. Therefore, the screening nurses of our medical record review committee evaluated 50 medical records for

a second time.

In **chapter five**, we investigated the internal reproducibility of the committee judgment. Therefore, 50 medical records were re-evaluated by the medical specialists of our medical record review committee.

In **chapter six**, we evaluated the external reproducibility of the committee judgment. Therefore, two committees from different hospitals evaluated each other's medical records.

Part III New developments

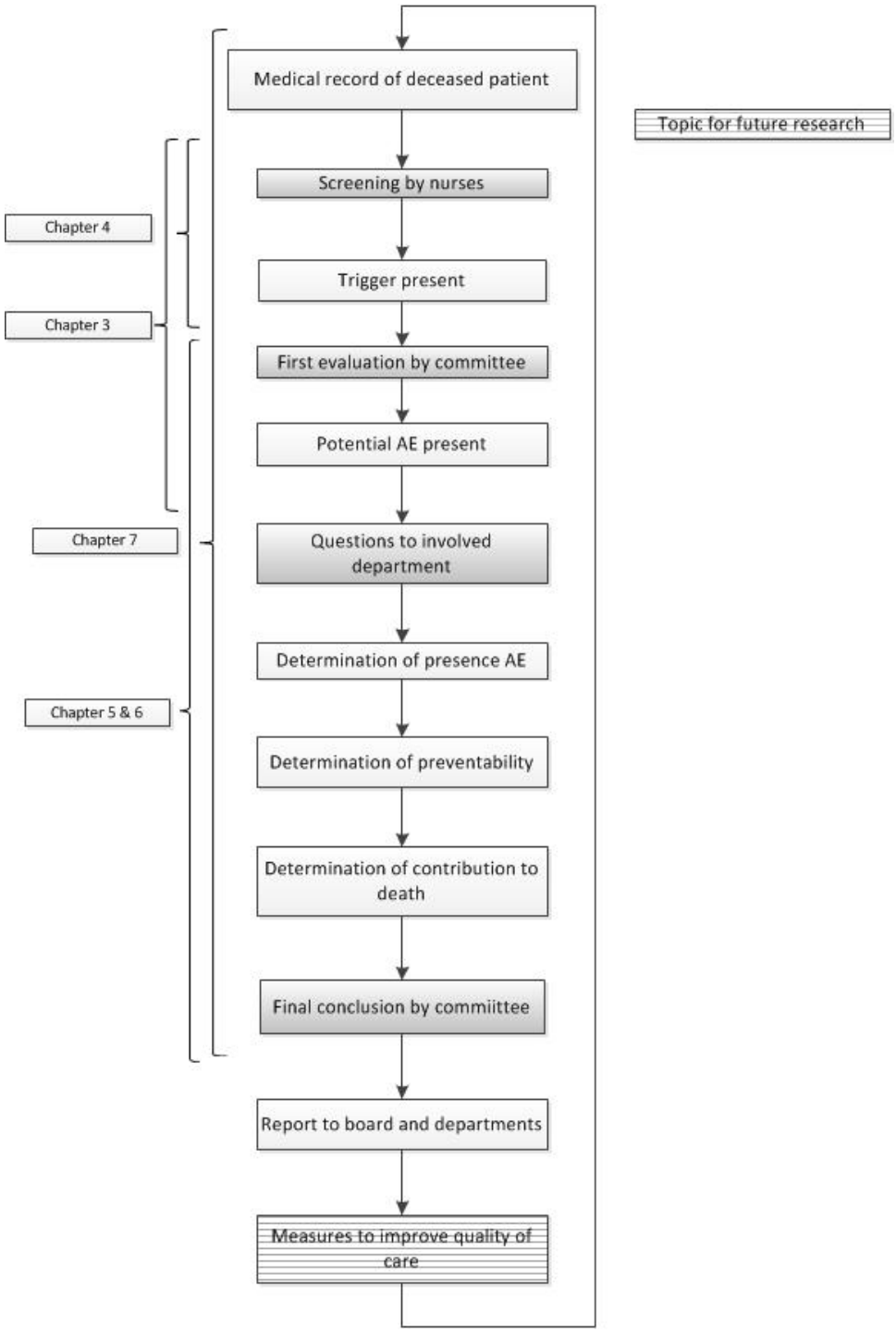
The results in part II made us decide to attempt to further improve the quality of the trigger tool. We discovered that still many cases are selected with the triggers, that don't contain an AE. This results in a waste of time and money.

In this part we assessed how the current methodology could be improved. Time- and cost wise, but also to improve its efficiency.

In **chapter seven**, we evaluated if the current trigger system could be improved by adding patient characteristics in the screening.

In **chapter eight**, we used a text mining tool for the screening of the medical records and compared these outcomes with the results of the manual screening by the nurses and specialists in the committee.

Figure 1: Overview of our studies concerning the process of improving quality and safety with medical record review



References

1.Europe WROf. Patient safety http://www.euro.who.int/en/health-topics/Health-systems/patient-safety/patient-safety.
2. de Vries EN, Ramrattan MA, Smorenburg SM, Gouma DJ, Boermeester MA. The incidence and nature of in-hospital adverse events: a systematic review. Qual Saf Health Care. 2008;17(3):216-23.
3. Najjar S, Hamdan M, Euwema MC, Vleugels A, Sermeus W, Massoud R, et al. The Global Trigger Tool shows that one out of seven patients suffers harm in Palestinian hospitals: challenges for launching a strategic safety plan. Int J Qual Health Care. 2013;25(6):640-7.
4. Aranaz-Andres JM, Aibar-Remon C, Vitaller-Murillo J, Ruiz-Lopez P, Limon-Ramirez R, Terol-Garcia E, et al. Incidence of adverse events related to health care in Spain: results of the Spanish National Study of Adverse Events. J Epidemiol Community Health. 2008;62(12):1022-9.
5. Baker GR, Norton PG, Flintoft V, Blais R, Brown A, Cox J, et al. The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. CMAJ. 2004;170(11):1678-86.
6. In: Kohn LT, Corrigan JM, Donaldson MS, editors. To Err is Human: Building a Safer Health System. Washington (DC)2000.
7. Silver MP, Hougland P, Elder S, Haug J, Pritchett T, Donnelly S, et al. Statewide identification of adverse events using retrospective nurse review: methods and outcomes. J Nurs Meas. 2007;15(3):220-32.
8. Hogan H, Healey F, Neale G, Thomson R, Vincent C, Black N. Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study. BMJ Qual Saf. 2012;21(9):737-45.
9. Kobewka DM, van Walraven C, Taljaard M, Ronksley P, Forster AJ. The prevalence of potentially preventable deaths in an acute care hospital: A retrospective cohort. Medicine (Baltimore). 2017;96(8):e6162.
10. Jha AK, Larizgoitia I, Audera-Lopez C, Prasopa-Plaizier N, Waters H, Bates DW. The global burden of unsafe medical care: analytic algorithmling of observational studies. BMJ Qual Saf. 2013;22(10):809-15.
11. Soop M, Fryksmark U, Koster M, Haglund B. The incidence of adverse events in Swedish hospitals: a retrospective medical record review study. Int J Qual Health Care. 2009;21(4):285-91.
12. Baines RJ, Langelaan M, de Bruijne MC, Wagner C. Is researching adverse events in hospital deaths a good way to describe patient safety in hospitals: a retrospective patient record review study. BMJ Open. 2015;5(7):e007380.
13. Zegers M, de Bruijne MC, Wagner C, Hoonhout LH, Waaijman R, Smits M, et al. Adverse events and potentially preventable deaths in Dutch hospitals: results of a retrospective patient record review study. Qual Saf Health Care. 2009;18(4):297-302.
14. van Gelderen SC, Zegers M, Boeijen W, Westert GP, Robben PB, Wollersheim HC. Evaluation of the organisation and effectiveness of internal audits to govern patient safety in hospitals: a mixed-methods study. BMJ Open. 2017;7(7):e015506.
15. Burnett S, Renz A, Wiig S, Fernandes A, Weggelaar AM, Calltorp J, et al. Prospects for comparing European hospitals in terms of quality and safety: lessons from a comparative study in five countries. Int J Qual Health Care. 2013;25(1):1-7.
16. de Feijter JM, de Grave WS, Muijtjens AM, Scherpbier AJJA, Koopmans RP. A Comprehensive Overview of Medical Error in Hospitals Using Incident-Reporting Systems, Patient Complaints and Chart Review of Inpatient Deaths. Plos One. 2012;7(2).
17. Hu Q, Wu B, Zhan M, Jia W, Huang Y, Xu T. Adverse events identified by the global trigger tool at a university hospital: A retrospective medical record review. J Evid Based Med. 2018.
18. Rafter N, Hickey A, Condell S, Conroy R, O'Connor P, Vaughan D, et al. Adverse events in healthcare: learning from mistakes. QJM. 2015;108(4):273-7.
19. Jochems A, Schouwenburg MG, Leeneman B, Franken MG, van den Eertwegh AJ, Haanen JB, et al. Dutch Melanoma Treatment Registry: Quality assurance in the care of patients with metastatic melanoma in the Netherlands. Eur J Cancer. 2017;72:156-65.
20. Hoeijmakers F, Beck N, Wouters M, Prins HA, Steup WH. National quality registries: how to improve the quality of data? J Thorac Dis. 2018;10(Suppl 29):S3490-S9.
21. Beck N, Hoeijmakers F, Wiegman EM, Smit HJM, Schramel FM, Steup WH, et al. Lessons learned from the Dutch Institute for Clinical Auditing: the Dutch algorithm for quality assurance in lung cancer treatment. J Thorac Dis. 2018;10(Suppl 29):S3472-S85.
22. Veen EJ, Janssen-Heijnen ML, Leenen LP, Roukema JA. The registration of complications in surgery: a learning curve. World J Surg. 2005;29(3):402-9.
23. Krol MW, De Boer D, Sixma H, Van Der Hoek L, Rademakers JJ, Delnoij DM. Patient experiences of inpatient hospital care: a department matter and a hospital matter. Int J Qual Health Care. 2015;27(1):17-25.
24. Kleefstra SM, Zandbelt LC, de Haes HJ, Kool RB. Trends in patient satisfaction in Dutch university medical

centers: room for improvement for all. BMC Health Serv Res. 2015;15:112.
25. Plexus K. Onderzoek kosten kwaliteitsmetingen. 2015.
26. Breyer JZ, Giacomazzi J, Kuhmmer R, Lima KM, Hammes LS, Ribeiro RA, et al. Hospital quality indicators: a systematic review. Int J Health Care Qual Assur. 2019;32(2):474-87.
27. Thomas JW, Hofer TP. Research evidence on the validity of risk-adjusted mortality rate as a measure of hospital quality of care. Med Care Res Rev. 1998;55(4):371-404.
28. Zhan C, Miller MR. Excess length of stay, charges, and mortality attributable to medical injuries during hospitalization. JAMA. 2003;290(14):1868-74.
29. Kehler DS, Kent D, Beaulac J, Strachan L, Wangasekara N, Chapman S, et al. Examining Patient Outcome Quality Indicators Based on Wait Time From Referral to Entry Into Cardiac Rehabilitation: A PILOT OBSERVATIONAL STUDY. J Cardiopulm Rehabil Prev. 2017;37(4):250-6.
30. Pincus D, Ravi B, Wasserstein D, Huang A, Paterson JM, Nathens AB, et al. Association Between Wait Time and 30-Day Mortality in Adults Undergoing Hip Fracture Surgery. JAMA. 2017;318(20):1994-2003.
31. Stephenson M, McArthur A, Giles K, Lockwood C, Aromataris E, Pearson A. Prevention of falls in acute hospital settings: a multi-site audit and best practice implementation project. Int J Qual Health Care. 2016;28(1):92-8.
32. Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. J Biomed Inform. 2003;36(1-2):131-43.
33. Jha AK, Kuperman GJ, Teich JM, Leape L, Shea B, Rittenberg E, et al. Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. J Am Med Inform Assoc. 1998;5(3):305-14.
34. Michel P, Quenon JL, de Sarasqueta AM, Scemama O. Comparison of three methods for estimating rates of adverse events and rates of preventable adverse events in acute care hospitals. BMJ. 2004;328(7433):199.
35. Thomas EJ, Petersen LA. Measuring errors and adverse events in health care. J Gen Intern Med. 2003;18(1):61-7.
36. Okoniewska B, Santana MJ, Holroyd-Leduc J, Flemons W, O'Beirne M, White D, et al. A framework to assess patient-reported adverse outcomes arising during hospitalization. BMC Health Serv Res. 2016;16(a):357.
37. Archer S, Hull L, Soukup T, Mayer E, Athanasiou T, Sevdalis N, et al. Development of a theoretical framework of factors affecting patient safety incident reporting: a theoretical review of the literature. BMJ Open. 2017;7(12):e017155.
38. Weingart SN, Callanan LD, Ship AN, Aronson MD. A physician-based voluntary reporting system for adverse events and medical errors. J Gen Intern Med. 2001;16(12):809-14.
39. Cifra CL, Jones KL, Ascenzi J, Bhalala US, Bembea MM, Fackler JC, et al. The morbidity and mortality conference as an adverse event surveillance tool in a paediatric intensive care unit. BMJ Qual Saf. 2014;23(11):930-8.
40. Vincent C, Neale G, Woloshynowych M. Adverse events in British hospitals: preliminary retrospective record review. BMJ. 2001;322(7285):517-9.
41. Woloshynowych M, Neale G, Vincent C. Case record review of adverse events: a new approach. Qual Saf Health Care. 2003;12(6):411-5.
42. Cihangir S, Borghans I, Hekkert K, Muller H, Westert G, Kool RB. A pilot study on record reviewing with a priori patient selection. BMJ Open. 2013;3(7).
43. Davis P, Lay-Yee R, Briant R, Ali W, Scott A, Schug S. Adverse events in New Zealand public hospitals I: occurrence and impact. N Z Med J. 2002;115(1167):U271.
44. Slappendel R. Meet dertigdagenmortaliteit en niet HSMR. Medisch Contact 2015:1452-5.
45. Van der Voort P dJE. Sterfte als maat voor kwaliteit Medisch Contact. 2007;62(43):1766-7.
46. van Gestel YR, Lemmens VE, Lingsma HF, de Hingh IH, Rutten HJ, Coebergh JW. The hospital standardized mortality ratio fallacy: a narrative review. Med Care. 2012;50(8):662-7.
47. Kahn JM, Kramer AA, Rubenfeld GD. Transferring critically ill patients out of hospital improves the standardized mortality ratio: a simulation study. Chest. 2007;131(1):68-75.
48. Hwang JI, Chin HJ, Chang YS. Characteristics associated with the occurrence of adverse events: a retrospective medical record review using the Global Trigger Tool in a fully digitalized tertiary teaching hospital in Korea. J Eval Clin Pract. 2014;20(1):27-35.
49. Nilsson L, Borgstedt-Risberg M, Soop M, Nylen U, Alenius C, Rutberg H. Incidence of adverse events in Sweden during 2013-2016: a cohort study describing the implementation of a national trigger tool. BMJ Open. 2018;8(3):e020833.
50. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. N Engl J Med. 1991;324(6):370-6.
51. Leape LL, Brennan TA, Laird N, Lawthers AG, Localio AR, Barnes BA, et al. The nature of adverse events in hospitalized patients. Results of the Harvard Medical Practice Study II. N Engl J Med. 1991;324(6):377-84.
52. Hiatt HH, Barnes BA, Brennan TA, Laird NM, Lawthers AG, Leape LL, et al. A study of medical injury and medi-

cal malpractice. N Engl J Med. 1989;321(7):480-4.
53. Unbeck M, Schildmeijer K, Henriksson P, Jurgensen U, Muren O, Nilsson L, et al. Is detection of adverse events affected by record review methodology? an evaluation of the "Harvard Medical Practice Study" method and the "Global Trigger Tool". Patient Saf Surg. 2013;7(1):10.
54. Hibbert PD, Molloy CJ, Hooper TD, Wiles LK, Runciman WB, Lachman P, et al. The application of the Global Trigger Tool: a systematic review. Int J Qual Health Care. 2016;28(6):640-9.
55. Wagner C. Onbedoelde schade in ziekenhuizen: resultaten dossieronderzoek naar patiëntveiligheid. Klachtenmanagement in de Zorg. 2007;4(3-4):28-31.
56. Baines R, Langelaan M, de Bruijne M, Spreeuwenberg P, Wagner C. How effective are patient safety initiatives? A retrospective patient record review study of changes to patient safety over time. BMJ Qual Saf. 2015;24(9):561-71.
57. Zegers M, de Bruijne MC, Wagner C, Groenewegen PP, Waaijman R, van der Wal G. Design of a retrospective patient record study on the occurrence of adverse events among patients in Dutch hospitals. BMC Health Serv Res. 2007;7:27.
58. Brennan TA, Leape LL. Adverse events, negligence in hospitalized patients: results from the Harvard Medical Practice Study. Perspect Healthc Risk Manage. 1991;11(2):2-8.
59. Wilson RM, Michel P, Olsen S, Gibberd RW, Vincent C, El-Assady R, et al. Patient safety in developing countries: retrospective estimation of scale and nature of harm to patients in hospital. BMJ. 2012;344:e832.
60. Schildmeijer K, Nilsson L, Arestedt K, Perk J. Assessment of adverse events in medical care: lack of consistency between experienced teams using the global trigger tool. BMJ Qual Saf. 2012;21(4):307-14.
61. Hofer TP, Bernstein SJ, DeMonner S, Hayward RA. Discussion between reviewers does not improve reliability of peer review of hospital quality. Med Care. 2000;38(2):152-61.
62. Sari AB, Sheldon TA, Cracknell A, Turnbull A. Sensitivity of routine system for reporting patient safety incidents in an NHS hospital: retrospective patient case note review. BMJ. 2007;334(7584):79.
63. Zegers M, de Bruijne MC, Wagner C, Groenewegen PP, van der Wal G, de Vet HC. The inter-rater agreement of retrospective assessments of adverse events does not improve with two reviewers per patient record. J Clin Epidemiol. 2010;63(1):94-102.
64. Sharek PJ, Parry G, Goldmann D, Bones K, Hackbarth A, Resar R, et al. Performance characteristics of a methodology to quantify adverse events over time in hospitalized patients. Health Serv Res. 2011;46(2):654-78.
65. Patient Safety: Rapid Assessment Methods for Estimating Hazards. WORLD HEALTH ORGANIZATION, Health DoHRf; 2003.

Part I

What is already known?

Chapter 2

A narrative systematic review of medical record analysis to improve patient safety in hospitals: is this method evidence based?

Klein DO, Rennenberg RJMW, Koopmans RP, Prins MH

Submitted

Abstract

Objectives: Evaluation of the two most used trigger tools according to the criteria of the WHO for evaluating methods.

Methods: We searched Embase, Pubmed and Cochrane databases for studies between 2000 and 2017. Studies were included if medical record review (MRR) was performed with either the Global Trigger Tool (GTT) or the Harvard Medical Practice Study (HMPS) in a hospital population. Quality assessment was performed in duplicate. Extracted outcome data included: prevalence of adverse events (AEs), kappa agreement for AEs, percentage agreement, summary of main results and costs of (detecting) AEs.

Results: 50 studies and several reports were included and results were reported for every criterion separately. MRR reveals more AEs than any other method. But at the same time, it detects different AEs compared to other methods. The costs of an AE were on average €4296. Considerable efforts have been made worldwide in healthcare to improve safety and to reduce errors. These efforts have resulted in some positive effects. The literature showed that MRR is focused on several domains of quality of care and seems suitable for both small as large cohorts. Furthermore, we found a moderate to substantial agreement for the presence of a trigger and a moderate to good agreement for the presence of an AE.

Conclusions: MRR with a trigger tool is a reasonably well researched method for the evaluation of the medical records for AEs. However, looking at the WHO criteria, much research is still lacking or of moderate quality. Especially for the cost of detecting AEs, valuable information is missing regarding. Moreover, knowledge of how MRR changes quality and safety of care should be evaluated.

Introduction

Several studies have shown significant rates of adverse events (AEs) that cause harm to patients during their stay in the hospital.[1-3] Therefore, interest in the implementation of quality and safety programs that prevent these harmful events has grown. According to the report 'to Err is Human' (1992) of the Institute of Medicine, at least 44,000 people (and possible even 98,000) died each year in US hospitals due to possibly preventable AEs.[3] An update, fifteen years after this first report, showed little improvement and stresses the need for further improving patient safety.[4] In the Netherlands, a report in 2009 by the Dutch institute for research in healthcare (NIVEL) evaluated care related harm in Dutch hospitals. It was estimated that yearly about 1,700 patients (4.1% of the total number of deaths in hospitals) die because of unintentional, but preventable, harm.[5] Follow-up showed slight improvement, but still 2.6% of the total number of inpatient deaths appeared to be preventable.[6] In other countries, an incidence between 2.5% and 11.5% has been found.[7-11]

Several factors contribute to the risk of AEs. First, the growing number of elderly patients with severe comorbidity[12,13] and, at the same time, increasing medical technical possibilities for more complex treatments and interventions. Second, due to this expansion in treatment options estimation of the possible health benefits is more complex. This could lead to treating more patients in which the predicted outcome is unclear, especially in the elderly with many comorbidities and vulnerable new-borns.[14-16] Furthermore, part-time working leads to more clinical hand-over moments and hence discontinuity of care which could result in AEs.[17-21] Teamwork and integrated care are necessary, but they depend on good communication within and between teams.[22,23] Also, economic limitations and cuts put health-care under pressure.[24-31] Together, these factors have serious impact on patient safety.

To diminish care related harm most hospitals have several instruments to detect preventable harm: direct observation, incident reporting systems, autopsy reports, mortality and morbidity conferences are some examples.[32-35]

For medical record review, trigger tools are often used to prevent time and cost consuming investigation of all records. The most well-known trigger systems are the Harvard Medical Practice Study trigger system (HMPS)[36] with 18 triggers[36,37] and the global trigger tool (GTT), developed by the Institute for Healthcare Improvement (IHI) with 54 triggers.[38] Information about sensitivity, specificity, positive and negative predictive values of these screening tools for detecting AEs is important in striving for effective and affordable methods to detect AEs. Also, MRR independently of the method used, is a costly process since it takes time of both experienced physicians and nurses hence it is of importance that the process is indeed improving quality and safety.

However, adequate detection of AEs is only a small part of the review process. The total procedure also involves feedback to the medical departments, adjustments in the delivery of care and hence improved outcome for patients resulting in less AEs. From this viewpoint, searching for AEs using medical record review is actually a screening method in which the AE is the disease or complication for which early intervention should improve the outcome (more of the same AEs). Previous studies mainly focused on the evaluation of a single part of this evaluation process.[39-41] We were interested in all stages of this screening method. Therefore, we investigated if the whole screening process based on medical

record review was evidence based according to the WHO criteria (box 1) for evaluating methods.[42]
With these WHO criteria in mind, we searched the literature for evidence about the use of this specific method to improve patient safety.

Box 1: WHO criteria for evaluating methods[42]

1. Effectiveness in capturing the extent of harm (in different environments)
2. Availability of reliable data
3. Suitability for large-scale or small, repeated studies
4. Costs (financial, human resources, time and burden on system)
5. Effectiveness in influencing policy
6. Effectiveness in influencing hospital and local safety procedures and outcomes
7. Synergy with other domains of quality of care

Methods

Search strategy and information sources
We identified potentially eligible studies by searching Pubmed, Embase and the Cochrane library for every criterion described in box 1. Our search was restricted to studies, in English or Dutch published between 2000 and 2017 because we assumed that older results might not be applicable because of rapidly changing healthcare. The search strategy and corresponding search terms are shown in the appendix, table 1. The flowchart for every criterion separately is shown in the appendix, table 2.

Selection criteria and process
Studies were included if, medical record review was performed with either the GTT or the HMPS in a hospital population with a wide variety of patient groups. Suitable study designs were observational studies (cross-sectional, retrospective or prospective cohorts, or case-control studies). Studies were excluded if they took place in a non-hospital setting, described only a single department, and used computer detection for finding triggers and/or AEs. Reviews, posters, comments, studies solely focusing on adverse drug events, patient populations <18 years were also excluded. Duplicate references were removed using the software program EndNote® (EndNote X8, Thomson Reuters, New York, USA). Afterwards, all retrieved citations based on the titles and abstracts were screened. Study selection was performed in duplicate by two independent reviewers (DK and RR) according to the aforementioned in and exclusion criteria. Study eligibility was thereafter assessed by reading the full text. Disagreements were resolved by discussion and consensus in which both reviewers had an equal vote.

Quality assessment
The quality of the included studies was evaluated using the Gilbert criteria complemented with criteria composed by Worster et al and Badcock et al.[43-45] The criteria from these three studies combined, resulted in 15 criteria on which every study was evaluated. Quality assessment was performed in duplicate by DK and RR. Rating categories used in the assessment were ''present'' or ''missing''. These were transformed into 1 and 0 and after scoring, the numbers were added together resulting in a score between 1 and 15. Scores

between 0-5 were seen as weak, between 6-10 as reasonable and between 11-15 as good. In table 4 (in the appendix) the results of this quality assessment are shown.

Data extraction and analysis

Data extraction was performed by DK. General data extracted from full-text included: author, year of publication, country, study aim, study design, instrument(s), sample characteristics, number of screening criteria, number of reviewers and number of records analysed for the measurement of the reliability. Extracted outcome data included: prevalence of AEs, kappa agreement for AEs, percentage agreement, summary of main results and costs of AEs (appendix, table 3). Using basic descriptive statistics (mean with confidence intervals if available) these variables were summarized. We report all results of the WHO criteria separately. Costs were not corrected for inflation. Different currencies were transformed to euros to make comparison easier. We used the exchange of August 2018 (1 euro = 1.13 US dollar). The results of this systematic review are presented according to the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines (appendix, table 5).[46]

Results

Our search resulted in a total of 832 citations (step 1-4), after title and abstract screening 752 were discarded leaving 80 full text articles to assess their eligibility. Another 63 were excluded after full text evaluation leaving 17 for inclusion. After reading the references in these studies we found 76 additional studies. The included studies originated from 24 countries and sample size varied from 96[47] to 210 million[48] patients. Some studies were relevant for more than 1 criterion. Below, we report the results for every criterion separately and according to the trigger tool (GTT or HMPS). The average quality of the included studies was reasonable (score 8.3).

Criterion 1: Effectiveness in capturing the extent of harm (in different environments)

16 studies were identified of reasonable quality, which compare the use of the medical record review method to the use of another method to detect AEs in a hospital setting. A wide diversity of methods were compared with the GTT such as reporting notification systems[49,50], a hospital survey on patient safety[51], patient safety indicators (PSIs)[49,52], complaints and claims by patients and their relative and incident reports[53,54]. For the HMPS, the outcomes were primarily compared with incident reports and furthermore patient-reported AEs55 and patient complaints.[56]

1.1 GTT

Kurutkan et al (2015) found that the GTT was 19 times more sensitive compared to internal voluntary reporting for the detection of AEs.[50] Another study by Farup et al (2015) detected an inverse association between the patient safety culture survey and AEs.[51] Kennerly et al (2014) found that voluntary reports and PSIs captured less than 5 percent of the total AEs[49] and this was also found by other studies.[52,57] Rutberg et al (2014) found that 6.3% of the AEs detected by the GTT where reported with voluntary reporting.[54] Mull et al (2015) compared three methods (veterans affair quality improvement programs, PSIs and voluntary incident reporting) with GTT and concluded that 12% of the AEs were also detected by one of the other methods.[53]

1.2 HMPS

Christiaans-Dingelhoff et al (2011) linked four reporting systems with HMPS reviewed records: informal and formal complaints by patients or relatives, medico-legal claims by patients or relatives and incident reports by healthcare professionals. Less than 4% of the AEs identified by record review were found in at least one of the four reporting systems.58

Michel et al (2004) compared three methods: cross sectional, prospective and retrospective review of records. The prospective and retrospective methods identified similar numbers of AEs (70% vs. 66% of the total) but the prospective method detected more preventable AEs (64% vs. 40%). The cross sectional method showed a large number of false positives and identified none of the most serious AEs.[35]

Several studies compared the number of AEs found when using incident reports compared to the use of HMPS. Blais et al (2008) showed that in 15.5% of the cases with an AE an incident report was present.[59] Also Sari detected underreporting by voluntary incident reporting, only 24% of all patient safety incidents and only 5% of those resulting in patient harm were detected with the HMPS method.[60] de Feijter compared both patient complaints, medical record review (MRR) and incident reporting systems, they found that the type of AE found, was depending on the method used. Therefore, they recommend using a combination of methods when assessing patient safety in a hospital.56 Macharias' (2016) findings were in line with the other studies.[61]

Weissmann et al (2008) found that 23% of the patients had at least 1 patient-reported AE and 11% according to medical record review. The agreement between these two were poor for occurrence of any type of AE and slightly better for life-threatening or serious AEs.[55] Bjertnaes et al (2015) showed a significant correlation between the two measurement methods.[62]

Criterion 2: Availability of reliable data

27 suitable studies of reasonable quality were found regarding the availability of reliable data in studies which a trigger tool method for medical record review.

2.1 Trigger tool

The IHI method showed a PPV of 30.4% (95%CI:13.3-47.6).[40,50,52,63-67] There was only one study that calculated the negative predictive value (NPV), which was 99%.[52] The agreement on the presence of a trigger had on average a moderate agreement (kappa= 0.48 (95%CI: 0.18-0.77).[2,63,64,66-68]

In studies using the HMPS procedure, the positive predictive value (PPV) was on average 33.4% (95% CI:21.1-45.8).[5,7-9,11,60,69-74] The agreement between the nurses on the presence of a trigger showed on average a substantial agreement (kappa=0.63; 95% CI 0.55-0.71).[5,7,60,70,71,73,75]

2.2 The AE assessment strategy

Within the studies using the IHI trigger tool[38] the kappa on the presence of an AE was on average 0.67 (good agreement, 95%CI:0.56-0.82).[1,2,40,50,57,63,64,66,68,76-79] In three studies the agreement on the severity of the AE was investigated, which showed an average K of 0.40 (fair agreement, 95%CI:0.06-0.73).[1,40,77,78]

In HMPS studies the kappa between medical doctors on the presence of an AE in the medical record was on average 0.58 (moderate agreement, 95%CI:0.33-0.82).[5,7,11,60,72,80-82]

Criterion 3: Suitability for large-scale or small, repeated studies
43 suitable studies of reasonable quality were found regarding the suitability of MRR for large-scale or small repeated studies.
In the last decades, several large scale studies have been executed to assess the prevalence of AEs in hospitals on a national level or the financial impact of AEs.[1,2,5,9,40,41,69,71,76,79,83-92] The smaller studies were used for the training of the reviewers[68], comparison with the detection rate of other methods [49,52,54,56,57,61,93-95], or for assessing the interrater reliability[72,96] of the review.

3.1 Size
The size of the studies varied from 15[68] (training) records to 40,851[87] and the number of hospitals investigated also varied, from 1 to 25 hospitals. Most of the studies were carried out in the US (16) and in Europe (14).

3.2 Cross country comparison
Deilkas et al[97] compared the AE rate between Norway and Sweden. There were significantly higher AE rates of surgical complications in Norwegian hospitals compared with Swedish hospitals. Swedish hospitals had significantly higher rates of pressure ulcers, falls and 'other' AEs. No significant difference between overall AE rates was found between the two countries.[97]

Criterion 4: Costs (financial, human resources, time and burden on system)
Of the 12 studies found concerning the costs of AEs, 6 reported these results based on the review of medical records, the quality of these 6 studies was reasonable. Other reported the costs based on a hospital claim database[48], a national medical and drugs claim database, hospital cost accounting system.[98] Furthermore, some studies used consensus based methodology, diagnostic coding error[99] and estimation of social cost.[100] The costs of an AE ranged between €2600 and €6436 with an average of €4296.[8,73,79,91,101,102] Next to the cost of an AE, the cost of finding an AE using MRR is of equal importance. No studies were available on the cost of detecting an AE. Based on our own data (unpublished) the cost of detecting an AE with HMPS was €150.000 on a yearly basis and approximately €1800 for a single potentially preventable AE.

Criterion 5: Effectiveness in influencing policy
We found no trials or studies but only reports of projects concerning this issue.

Short overview of studies mentioned in the context of PASQ (European Union Network for Patient Safety and Quality of Care)
In the Netherlands, the reports by NIVEL have been published in the context of a research program named patient safety in the Netherlands. The first study was first performed to gain insight regarding AEs in Dutch hospitals.[5] The two follow-up studies were executed to evaluate whether the safety programs had a positive influence on these AEs.[85,103]

According to experience in Norway and Sweden, MRR by the GTT method gives a valua-

ble overview of kind and incidence of AEs affecting patients and a good starting point for intensified patient safety improvement work.[104,105]

Criterion 6: Effectiveness in influencing hospital and local safety procedures and outcomes

Our search revealed 5 studies investigating changes in AE rates during the study period. Kennerly et al (2013) showed a 7% reduction in AEs in 2 years (on average 3.5% per year).[88] Suarez et al (2014) found during a 6-year study period a decrease of 2.5% (on average 0,4% per year).[106] Deilkas et al (2017) showed that AEs rates decreased from 16.1% to 13.0% in two years (1.55% per year).[87]

However, Rutberg et al (2014) found no improvement during the 4 year study period in which the GTT was used, despite several initiatives for improving the quality in the hospital.[107] Same for Landrigan (2010)[1] and Mortaro (2017)[90]. Three national studies in the Netherlands (2004-2012) showed no changes in overall AE but did show a decrease of 45% regarding preventable AEs.[103] However, the latest update in 2017 didn't show a further decrease of the preventable AEs and preventable deaths but did show a 2% decrease in overall AE.[108] Landrigan et al also found slight improvements in a 5 year time period in the US.[1]

Criterion 7: Synergy with other domains of quality of care

We found no trials or studies but only reports of projects concerning this issue. The IHI has defined six domains of quality of care.[109] Medical record review has common ground with a few of these domains; safe, effective, timely and equitable. During medical record review a committee assesses whether AEs have occurred. The goal is to improve care, making it safer. Furthermore, it is evaluated whether the specific treatment of a particular patient was correct, right on time (effective and timely) and independent of personal characteristics (equitable). The project 'Deepening our Understanding of Quality Improvement in Europe' (DUQUE) investigated the relation between quality systems and patient related outcomes. Almost 200 hospitals in 8 European countries participated in DUQUE. Beside questionnaires, also medical record review and data registries were analysed. One of the conclusions of this project was that presence of quality systems has a positive effect on the safety culture in a hospital.[109]

Discussion

Our study clearly shows there is abundant literature concerning MRR in hospitals. However, almost 75% of these articles were of rather moderate quality (step 1-4). The first four WHO-criteria relate to characteristics of medical record review concerning validity, reliability and costs. They could effortlessly be extracted from the existing literature. The last three criteria relate to the ability of MRR to generate improvements in safety procedures and the quality of safety programs. Data concerning criterion five and seven was indirectly described or concealed in reports and therefore we were unable to evaluate the quality according to the quality checklist.

The literature we found in relation to the first criterion showed that MRR reveals more AEs than any other method. But at the same time, MRR detects different AEs compared to

other methods. Furthermore, we found a moderate agreement for the presence of a trigger and a good agreement for the presence of an AE for the GTT. For the HMPS we found a substantial agreement for the presence of a trigger and a moderate agreement for the presence of an AE. Also, MRR seems suitable for both small and large cohorts as shown in several studies with different sample sizes.

The costs concerning AEs can be the cost of the event itself (which is usually the topic of the literature we found), but also the cost of MRR. It is striking that most studies investigating costs of AEs only evaluated costs related to the event. The only study we found evaluating also the costs of the detection method was published by Bates et al in 1995, but was not included in the current study because we selected studies published from the year 2000 and onwards. The costs for the detection of a single AE in 1995 was 103€ and 241€ for a preventable AE (€11.10 for every admission). Translating this to the current situation means a considerable amount of money for the detection instrument, let alone the costs of the AEs themselves. Because there is no agreement on which costs exactly should be taken into account concerning AEs, comparison between studies is difficult. For future research it is important to have a complete overview of all costs involved. It should contain the cost of the detection instrument, the direct cause of the AE itself, but also loss of working days, up to implementation of other protocols and their costs to prevent AEs. Only with these total costs we will be able to estimate the costs per quality adjusted life year to see if this is acceptable. The above is also underlined by a report of Øvretveit (2009)[110] and a more recent report by the Organization for Economic Co-operation and Development. The results of the fifth criterion show that MRR can have an effect on healthcare policy. In Europe, the network for patient safety and quality of care (PASQ) has been active for 4 years. This network was co-funded and supported by the European Commission within the Public Health Program. Its focus was 'to improve Patient Safety and Quality of Care in Europe by supporting the implementation of good organizational practices and safe clinical practices in health care organizations and through sharing of information and experiences'. PASQ builds on the experience of the European Union Network for Patient Safety (EUNet-PaS) project (2008-2010) which established patient safety platforms in several European member states. The main outcome will be the consolidation of the permanent network for patient safety.

Considerable efforts have been made worldwide in the healthcare systems to improve safety and to reduce errors in the treatment of patients. As is shown in criterion 6, these efforts have translated into only slight improvements in the overall safety of patients or in better quality of care. Finally, the last criterion shows that medical record review has focused on several domains of quality of care.

Michel has composed an overview of strength and weaknesses of available methods for assessing AEs.[111] MRR is one of these methods. Since then (2003) no update has been performed to investigate how well the trigger tools comply according to the WHO criteria. Instead of giving an overview of the available methods, we decided to focus on the 2 most used trigger tools. Also, in this overview Michel didn't give an insight into the quality of the included studies, rather compared the methods with each other. Furthermore, he evaluated the evidence-based rating of the methods for estimating AEs, for each criterion.

Our systematic review had several strengths and some limitations. We used extensive search criteria (table S1) in several databases and also screened the references of the finally selected articles which, to our opinion, minimize the risk of missing important literature. Moreover, we used an accepted WHO strategy to evaluate this screening tool. Furthermore, we combined several quality checklists to evaluate the quality of the included studies.

An important limitation for drawing aggregate conclusions was the different methods studies deployed. Although the same trigger tool method was used, almost every study adapted the triggers or the review process slightly. Besides that, different definitions and scales were used for both AEs and their preventability. Moreover, some studies used external reviewers, some internal reviewers or experienced reviewers. Although the intention was to evaluate the quality of all included studies, this was only possible for five out of seven WHO criteria. The reason was that the others were not suitable for evaluation according to our fifteen quality requirements. These were actually more reports than studies. For other studies, we found that important detailed information was not reported (for example: information on positive and negative agreement, reproducibility regarding the individual triggers). We only searched for studies from the year 2000 and onwards. There is a chance that important older studies have therefore been missed. However, we doubt that these findings would still be generalizable to today's care because of rapidly developing healthcare and quality and safety improvements. Also, we were forced to write a systematic narrative review, due to the number of different studies we were not able to combine the numbers and create a meta-analysis.

In conclusion, medical record review with either the GTT or the HMPS trigger tool is a reasonably well researched method for the evaluation of the medical records for AEs. However, looking at the WHO criteria for the evaluation of methods, much research is still lacking or of moderate quality. Also, more information is needed concerning costs of the detection method and the improvements in care and patient outcome. Only with this information MRR could be evaluated on its cost-effectiveness. Moreover, more insight in how MRR changes quality and safety of care is needed. We found no studies analysing the whole string that starts with the application of the triggers and ends with quality and safety improvement for individual patients at acceptable costs.

References

1. Landrigan CP, Parry GJ, Bones CB, Hackbarth AD, Goldmann DA, Sharek PJ. Temporal trends in rates of patient harm resulting from medical care. The New England journal of medicine. 2010;363(22):2124-34.
2. Najjar S, Hamdan M, Euwema MC, Vleugels A, Sermeus W, Massoud R, et al. The Global Trigger Tool shows that one out of seven patients suffers harm in Palestinian hospitals: challenges for launching a strategic safety plan. International journal for quality in health care : journal of the International Society for Quality in Health Care. 2013;25(6):640-7.
3. Kohn L, T., Corrigan J.M., Donaldson, M. To Err is human: building a safer health system. Washington, DC: 1999.
4. Foundation NPS. Free from Harm: Accelerating Patient Safety Improvement Fifteen Years after To Err Is Human 2015.
5. Zegers M, de Bruijne MC, Wagner C, Hoonhout LH, Waaijman R, Smits M, et al. Adverse events and potentially preventable deaths in Dutch hospitals: results of a retrospective patient record review study. Qual Saf Health Care. 2009;18(4):297-302.
6. M. L, Broekens MA, de Bruijne MC, de Groot JF, Moesker MJ, Porte PJ, et al. Monitor Zorggerelateerde Schade 2015/2016. NIVEL, 2017.
7. Baker GR, Norton PG, Flintoft V, Blais R, Brown A, Cox J, et al. The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2004;170(11):1678-86.
8. Vincent C, Neale G, Woloshynowych M. Adverse events in British hospitals: preliminary retrospective record review. BMJ. 2001;322(7285):517-9.
9. Davis P, Lay-Yee R, Briant R, Ali W, Scott A, Schug S. Adverse events in New Zealand public hospitals I: occurrence and impact. N Z Med J. 2002;115(1167):U271.
10. Wilson RM, Runciman WB, Gibberd RW, Harrison BT, Newby L, Hamilton JD. The Quality in Australian Health Care Study. The Medical journal of Australia. 1995;163(9):458-71.
11. Thomas EJ, Studdert DM, Burstin HR, Orav EJ, Zeena T, Williams EJ, et al. Incidence and types of adverse events and negligent care in Utah and Colorado. Medical care. 2000;38(3):261-71.
12. Tran J, Norton R, Conrad N, Rahimian F, Canoy D, Nazarzadeh M, et al. Patterns and temporal trends of comorbidity among adult patients with incident cardiovascular disease in the UK between 2000 and 2014: A population-based cohort study. PLoS medicine. 2018;15(3):e1002513.
13. Piccirillo JF, Vlahiotis A, Barrett LB, Flood KL, Spitznagel EL, Steyerberg EW. The changing prevalence of comorbidity across the age spectrum. Crit Rev Oncol Hematol. 2008;67(2):124-32.
14. Hamaker ME, Acampo T, Remijn JA, van Tuyl SA, Pronk A, van der Zaag ES, et al. Diagnostic choices and clinical outcomes in octogenarians and nonagenarians with iron-deficiency anemia in the Netherlands. Journal of the American Geriatrics Society. 2013;61(4):495-501.
15. Glass HC, Costarino AT, Stayer SA, Brett CM, Cladis F, Davis PJ. Outcomes for extremely premature infants. Anesthesia and analgesia. 2015;120(6):1337-51.
16. Ashby M. Caring for dying patients is not about prolonging life at all costs. BMJ (Clinical research ed). 2013;346:f3027.
17. Laine C, Goldman L, Soukup JR, Hayes JG. The impact of a regulation restricting medical house staff working hours on the quality of patient care. Jama. 1993;269(3):374-8.
18. Petersen LA, Brennan TA, O'Neil AC, Cook EF, Lee TH. Does housestaff discontinuity of care increase the risk for preventable adverse events? Annals of internal medicine. 1994;121(11):866-72.
19. Eggins S, Slade D. Communication in Clinical Handover: Improving the Safety and Quality of the Patient Experience. J Public Health Res. 2015;4(3):666.
20. de Jong JD, Heiligers P, Groenewegen PP, Hingstman L. Part-time and full-time medical specialists, are there differences in allocation of time? BMC health services research. 2006;6:26.
21. Arora V, Johnson J, Lovinger D, Humphrey HJ, Meltzer DO. Communication failures in patient sign-out and suggestions for improvement: a critical incident analysis. Quality & safety in health care. 2005;14(6):401-7.
22. Birk K, Paden L, Markic M. Adverse event reporting in Slovenia - the influence of safety culture, supervisors and communication. Vojnosanit Pregl. 2016;73(8):714-22.
23. Wami SD, Demssie AF, Wassie MM, Ahmed AN. Patient safety culture and associated factors: A quantitative and qualitative study of healthcare workers' view in Jimma zone Hospitals, Southwest Ethiopia. BMC health services research. 2016;16:495.
24. Aranaz-Andres JM, Aibar-Remon C, Vitaller-Murillo J, Ruiz-Lopez P, Limon-Ramirez R, Terol-Garcia E, et al. Incidence of adverse events related to health care in Spain: results of the Spanish National Study of Adverse Events. J Epidemiol Community Health. 2008;62(12):1022-9.
25. Leonard M, Graham S, Bonacum D. The human factor: the critical importance of effective teamwork and communication in providing safe care. Quality & safety in health care. 2004;13 Suppl 1:i85-90.

26. Lingard L, Espin S, Whyte S, Regehr G, Baker GR, Reznick R, et al. Communication failures in the operating room: an observational classification of recurrent types and effects. Quality & safety in health care. 2004;13(5):330-4.
27. Vincent C, Taylor-Adams S, Stanhope N. Framework for analysing risk and safety in clinical medicine. BMJ (Clinical research ed). 1998;316(7138):1154-7.
28. Kang JH, Kim CW, Lee SY. Nurse-Perceived Patient Adverse Events depend on Nursing Workload. Osong public health and research perspectives. 2016;7(1):56-62.
29. Rogers AE, Hwang WT, Scott LD, Aiken LH, Dinges DF. The working hours of hospital staff nurses and patient safety. Health affairs (Project Hope). 2004;23(4):202-12.
30. Cho SH, Ketefian S, Barkauskas VH, Smith DG. The effects of nurse staffing on adverse events, morbidity, mortality, and medical costs. Nurs Res. 2003;52(2):71-9.
31. Olds DM, Clarke SP. The effect of work hours on adverse events and errors in health care. Journal of safety research. 2010;41(2):153-62.
32. Forster AJ, Andrade J, Van Walraven C. Validation of a discharge summary term search method to detect adverse events. Journal of the American Medical Informatics Association. 2005;12(2):200-6.
33. Forster AJ, Worthington JR, Hawken S, Bourke M, Rubens F, Shojania K, et al. Using prospective clinical surveillance to identify adverse events in hospital. BMJ quality & safety. 2011;20(9):756-63.
34. Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. J Biomed Inform. 2003;36(1-2):131-43.
35. Michel P, Quenon JL, de Sarasqueta AM, Scemama O. Comparison of three methods for estimating rates of adverse events and rates of preventable adverse events in acute care hospitals. BMJ. 2004;328(7433):199.
36. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. N Engl J Med. 1991;324(6):370-6.
37. Brennan TA, Leape LL. Adverse events, negligence in hospitalized patients: results from the Harvard Medical Practice Study. Perspect Healthc Risk Manage. 1991;11(2):2-8.
38. Griffin FA. IHI Global Trigger Tool for Measuring Adverse Events (Second Edition) IHI Innovation Series white paper. Cambridge, Massachusetts: Institute for Healthcare Improvement. 2009.
39. Ock M, Lee SI, Jo MW, Lee JY, Kim SH. Assessing reliability of medical record reviews for the detection of hospital adverse events. Journal of Preventive Medicine and Public Health. 2015;48(5):239-48.
40. Sharek PJ, Parry G, Goldmann D, Bones K, Hackbarth A, Resar R, et al. Performance characteristics of a methodology to quantify adverse events over time in hospitalized patients. Health Serv Res. 2011;46(2):654-78.
41. Hwang JI, Chin HJ, Chang YS. Characteristics associated with the occurrence of adverse events: A retrospective medical record review using the Global Trigger Tool in a fully digitalized tertiary teaching hospital in Korea. Journal of evaluation in clinical practice. 2014;20(1):27-35.
42. Patient Safety: Rapid Assessment Methods for Estimating Hazards. WORLD HEALTH ORGANIZATION, Health DoHRf; 2003.
43. Gilbert EH, Lowenstein SR, Koziol-McLain J, Barta DC, Steiner J. Chart reviews in emergency medicine research: Where are the methods? Annals of emergency medicine. 1996;27(3):305-8.
44. Badcock D, Kelly AM, Kerr D, Reade T. The quality of medical record review studies in the international emergency medicine literature. Annals of emergency medicine. 2005;45(4):444-7.
45. Worster A, Bledsoe RD, Cleve P, Fernandes CM, Upadhye S, Eva K. Reassessing the methods of medical record review studies in emergency medicine research. Annals of emergency medicine. 2005;45(4):448-51.
46. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Journal of clinical epidemiology. 2009;62(10):e1-34.
47. Ock M, Lee SI, Jo MW, Lee JY, Kim SH. Assessing Reliability of Medical Record Reviews for the Detection of Hospital Adverse Events. J Prev Med Public Health. 2015;48(5):239-48.
48. David G, Gunnarsson CL, Waters HC, Horblyuk R, Kaplan HS. Economic measurement of medical errors using a hospital claims database. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2013;16(2):305-10.
49. Kennerly DA, Kudyakov R, da Graca B, Saldana M, Compton J, Nicewander D, et al. Characterization of adverse events detected in a large health care delivery system using an enhanced global trigger tool over a five-year interval. Health services research. 2014;49(5):1407-25.
50. Kurutkan MN, Usta E, Orhan F, Simsekler MC. Application of the IHI Global Trigger Tool in measuring the adverse event rate in a Turkish healthcare setting. The International journal of risk & safety in medicine. 2015;27(1):11-21.
51. Farup PG. Are measurements of patient safety culture and adverse events valid and reliable? Results from a cross sectional study. BMC health services research. 2015;15:186.

52. Naessens JM, Campbell CR, Huddleston JM, Berg BP, Lefante JJ, Williams AR, et al. A comparison of hospital adverse events identified by three widely used detection methods. Int J Qual Health Care. 2009;21(4):301-7.
53. Mull HJ, Brennan CW, Folkes T, Hermos J, Chan J, Rosen AK, et al. Identifying Previously Undetected Harm: Piloting the Institute for Healthcare Improvement's Global Trigger Tool in the Veterans Health Administration. Quality management in health care. 2015;24(3):140-6.
54. Rutberg H, Risberg MB, Sjodahl R, Nordqvist P, Valter L, Nilsson L. Characterisations of adverse events detected in a university hospital: A 4-year study using the Global Trigger Tool method. BMJ open. 2014;4 (5) (no pagination)(e004879).
55. Weissman JS, Schneider EC, Weingart SN, Epstein AM, David-Kasdan J, Feibelmann S, et al. Comparing patient-reported hospital adverse events with medical record review: do patients know something that hospitals do not? Annals of internal medicine. 2008;149(2):100-8.
56. de Feijter JM, de Grave WS, Muijtjens AM, Scherpbier AJ, Koopmans RP. A comprehensive overview of medical error in hospitals using incident-reporting systems, patient complaints and chart review of inpatient deaths. PloS one. 2012;7(2):e31125.
57. Classen DC, Resar R, Griffin F, Federico F, Frankel T, Kimmel N, et al. 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. Health affairs (Project Hope). 2011;30(4):581-9.
58. Christiaans-Dingelhoff I, Smits M, Zwaan L, Lubberding S, van der Wal G, Wagner C. To what extent are adverse events found in patient records reported by patients and healthcare professionals via complaints, claims and incident reports? BMC Health Serv Res. 2011;11:49.
59. Blais RPB, Derson MD, MSc*; Bartlett, Gillian PhD†; Tamblyn, Robyn PhD‡§. Can We Use Incident Reports to Detect Hospital Adverse Events? Journal of patient safety. 2008;4(1):9-12.
60. Sari AB, Sheldon TA, Cracknell A, Turnbull A. Sensitivity of routine system for reporting patient safety incidents in an NHS hospital: retrospective patient case note review. BMJ. 2007;334(7584):79.
61. Macharia WM, Muteshi CM, Wanyonyi SZ, Mukaindo AM, Ismail A, Ekea H, et al. Comparison of the prevalence and characteristics of inpatient adverse events using medical records review and incident reporting. S Afr Med J. 2016;106(10):1021-36.
62. Bjertnaes O, Deilkas ET, Skudal KE, Iversen HH, Bjerkan AM. The association between patient-reported incidents in hospitals and estimated rates of patient harm. International journal for quality in health care : journal of the International Society for Quality in Health Care. 2015;27(1):26-30.
63. Schildmeijer K, Nilsson L, Arestedt K, Perk J. Assessment of adverse events in medical care: lack of consistency between experienced teams using the global trigger tool. BMJ Qual Saf. 2012;21(4):307-14.
64. Kennerly DA, Saldana M, Kudyakov R, da Graca B, Nicewander D, Compton J. Description and evaluation of adaptations to the global trigger tool to enhance value to adverse event reduction efforts. Journal of patient safety. 2013;9(2):87-95.
65. Unbeck M, Schildmeijer K, Henriksson P, Jurgensen U, Muren O, Nilsson L, et al. Is detection of adverse events affected by record review methodology? an evaluation of the "Harvard Medical Practice Study" method and the "Global Trigger Tool". Patient Saf Surg. 2013;7(1):10.
66. Hwang JI, Chin HJ, Chang YS. Characteristics associated with the occurrence of adverse events: a retrospective medical record review using the Global Trigger Tool in a fully digitalized tertiary teaching hospital in Korea. Journal of evaluation in clinical practice. 2014;20(1):27-35.
67. Naessens JM, O'Byrne TJ, Johnson MG, Vansuch MB, McGlone CM, Huddleston JM. Measuring hospital adverse events: Assessing inter-rater reliability and trigger performance of the Global Trigger Tool. International Journal for Quality in Health Care. 2010;22(4):266-74.
68. Classen. Development and Evaluation of the Institute for Healthcare Improvement Global Trigger Tool. Journal of patient safety. 2008;4(3):169-77.
69. Akbari Sari A, Doshmangir L, Torabi F, Rashidian A, Sedaghat M, Ghomi R, et al. The Incidence, Nature and Consequences of Adverse Events in Iranian Hospitals. Arch Iran Med. 2015;18(12):811-5.
70. Thomas EJ, Lipsitz SR, Studdert DM, Brennan TA. The reliability of medical record review for estimating adverse event rates. Annals of internal medicine. 2002;136(11):812-6.
71. Soop M, Fryksmark U, Koster M, Haglund B. The incidence of adverse events in Swedish hospitals: a retrospective medical record review study. International journal for quality in health care : journal of the International Society for Quality in Health Care. 2009;21(4):285-91.
72. Zegers M, de Bruijne MC, Wagner C, Groenewegen PP, van der Wal G, de Vet HC. The inter-rater agreement of retrospective assessments of adverse events does not improve with two reviewers per patient record. Journal of clinical epidemiology. 2010;63(1):94-102.
73. Sousa P, Uva AS, Serranheira F, Nunes C, Leite ES. Estimating the incidence of adverse events in Portuguese hospitals: a contribution to improving quality and patient safety. BMC health services research. 2014;14:311.
74. Klein DO, Rennenberg R, Koopmans RP, Prins MH. The ability of triggers to retrospectively predict potentially

preventable adverse events in a sample of deceased patients. Prev Med Rep. 2017;8:250-5.
75. Wilson RM, Michel P, Olsen S, Gibberd RW, Vincent C, El-Assady R, et al. Patient safety in developing countries: retrospective estimation of scale and nature of harm to patients in hospital. BMJ (Clinical research ed). 2012;344:e832.
76. Asavaroengchai S, Sriratanaban J, Hiransuthikul N, Supachutikul A. Identifying adverse events in hospitalized patients using Global Trigger Tool in Thailand. Asian Biomedicine. 2009;3(5):545-50.
77. O'Leary KJ, Devisetty VK, Patel AR, Malkenson D, Sama P, Thompson WK, et al. Comparison of traditional trigger tool to data warehouse based screening for identifying hospital adverse events. BMJ quality & safety. 2013;22(2):130-8.
78. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.
79. Adler L, Yi D, Li M, McBroom B, Hauck L, Sammer C, et al. Impact of Inpatient Harms on Hospital Finances and Patient Clinical Outcomes. Journal of patient safety. 2018;14(2):67-73.
80. Soop M, Fryksmark U, Koster M, Haglund B. The incidence of adverse events in Swedish hospitals: A retrospective medical record review study. International Journal for Quality in Health Care. 2009;21(4):285-91.
81. Naessens JM, O'Byrne TJ, Johnson MG, Vansuch MB, McGlone CM, Huddleston JM. Measuring hospital adverse events: assessing inter-rater reliability and trigger performance of the Global Trigger Tool. International journal for quality in health care : journal of the International Society for Quality in Health Care. 2010;22(4):266-74.
82. Zwaan L, De Bruijne M, Wagner C, Thijs A, Smits M, Van Der Wal G, et al. Patient record review of the incidence, consequences, and causes of diagnostic adverse events. Archives of internal medicine. 2010;170(12):1015-21.
83. Good VS, Saldana M, Gilder R, Nicewander D, Kennerly DA. Large-scale deployment of the Global Trigger Tool across a large hospital system: refinements for the characterisation of adverse events to support patient safety learning opportunities. BMJ quality & safety. 2011;20(1):25-30.
84. Garrett PR, Jr., Sammer C, Nelson A, Paisley KA, Jones C, Shapiro E, et al. Developing and implementing a standardized process for global trigger tool application across a large health system. Joint Commission journal on quality and patient safety. 2013;39(7):292-7.
85. Baines RJ, Langelaan M, de Bruijne MC, Asscheman H, Spreeuwenberg P, van de Steeg L, et al. Changes in adverse event rates in hospitals over time: a longitudinal retrospective patient record review study. BMJ quality & safety. 2013;22(4):290-8.
86. Baines RJ, de Bruijne MC, Langelaan M, Wagner C. What are the safety risks for patients undergoing treatment by multiple specialties: a retrospective patient record review study. BMC health services research. 2013;13:497.
87. Deilkas ET, Bukholm G, Lindstrom JC, Haugen M. Monitoring adverse events in Norwegian hospitals from 2010 to 2013. BMJ open. 2015;5(12):e008576.
88. Kennerly DA, Saldana M, Kudyakov R, da Graca B, Nicewander D, Compton J. Description and evaluation of adaptations to the global trigger tool to enhance value to adverse event reduction efforts. Journal of patient safety. 2013;9(2):87-95.
89. Kurutkan MN, Usta E, Orhan F, Simsekler MCE. Application of the IHI Global Trigger Tool in measuring the adverse event rate in a Turkish healthcare setting. International Journal of Risk and Safety in Medicine. 2015;27(1):11-21.
90. Mortaro A, Moretti F, Pascu D, Tessari L, Tardivo S, Pancheri S, et al. Adverse Events Detection Through Global Trigger Tool Methodology: Results From a 5-Year Study in an Italian Hospital and Opportunities to Improve Inter-rater Reliability. J Patient Saf. 2017.
91. Rafter N, Hickey A, Conroy RM, Condell S, O'Connor P, Vaughan D, et al. The Irish National Adverse Events Study (INAES): the frequency and nature of adverse events in Irish hospitals-a retrospective record review study. BMJ quality & safety. 2017;26(2):111-9.
92. Von Plessen C, Kodal AM, Anhoj J. Experiences with global trigger tool reviews in five Danish hospitals: An implementation study. BMJ open. 2012;2 (5) (no pagination)(e001324).
93. Kobayashi M, Ikeda S, Kitazawa N, Sakai H. Validity of retrospective review of medical records as a means of identifying adverse events: comparison between medical records and accident reports. Journal of evaluation in clinical practice. 2008;14(1):126-30.
94. Christiaans-Dingelhoff I, Smits M, Zwaan L, Lubberding S, van der Wal G, Wagner C. To what extent are adverse events found in patient records reported by patients and healthcare professionals via complaints, claims and incident reports? BMC health services research. 2011;11:49.
95. Weissman JS, Schneider EC, Weingart SN, Epstein AM, David-Kasdan J, Feibelmann S, et al. Comparing patient-reported hospital adverse events with medical record review: Do patients know something that hospitals do not? Annals of internal medicine. 2008;149(2):100-8.
96. Schildmeijer K, Nilsson L, Arestedt K, Perk J. Assessment of adverse events in medical care: Lack of consistency between experienced teams using the global trigger tool. BMJ Quality and Safety. 2012;21(4):307-14.

97. Deilkas ET, Risberg MB, Haugen M, Lindstrom JC, Nylen U, Rutberg H, et al. Exploring similarities and differences in hospital adverse event rates between Norway and Sweden using Global Trigger Tool. BMJ open. 2017;7(3):e012492.
98. Pappas SH. The cost of nurse-sensitive adverse events. J Nurs Adm. 2008;38(5):230-6.
99. Wardle G, Wodchis WP, Laporte A, Anderson GM, Ross Baker G. The sensitivity of adverse event cost estimates to diagnostic coding error. Health services research. 2012;47(3 Pt 1):984-1007.
100. Goodman JC, Villarreal P, Jones B. The social cost of adverse medical events, and what we can do about it. Health affairs (Project Hope). 2011;30(4):590-5.
101. Brown P, McArthur C, Newby L, Lay-Yee R, Davis P, Briant R. Cost of medical injury in New Zealand: a retrospective cohort study. J Health Serv Res Policy. 2002;7 Suppl 1:S29-34.
102. Hoogervorst-Schilp J, Langelaan M, Spreeuwenberg P, de Bruijne MC, Wagner C. Excess length of stay and economic consequences of adverse events in Dutch hospital patients. BMC health services research. 2015;15:531.
103. Baines R, Langelaan M, de Bruijne M, Spreeuwenberg P, Wagner C. How effective are patient safety initiatives? A retrospective patient record review study of changes to patient safety over time. BMJ quality & safety. 2015;24(9):561-71.
104. Deilkas ET, Bukholm G, Lindstrom JC, Haugen M. Monitoring adverse events in Norwegian hospitals from 2010 to 2013. BMJ open. 2015;5 (12) (no pagination)(e008576).
105. Best practices in patient safety Federal Ministry of Health - Germany 2017.
106. Suarez C, Menendez MD, Alonso J, Castano N, Alonso M, Vazquez F. Detection of adverse events in an acute geriatric hospital over a 6-year period using the global trigger tool. Journal of the American Geriatrics Society. 2014;62(5):896-900.
107. Rutberg H, Borgstedt Risberg M, Sjodahl R, Nordqvist P, Valter L, Nilsson L. Characterisations of adverse events detected in a university hospital: a 4-year study using the Global Trigger Tool method. Bmj Open. 2014;4(5):e004879.
108. Langelaan M, Broekens M, de Bruijne M, de Groot J, Moesker M, Porte P, et al. Monitor Zorggerelateerde Schade 2015/2016 - Dossieronderzoek bij overleden patiënten in Nederlandse ziekenhuizen. NIVEL 2017.
109. Across the Chasm: Six Aims for Changing the Health Care System. Institue for Healthcare Improvement.
110. Øvretveit J. Does improving quality save money? A review of evidenve of which improbement to quality reduce costs to health service providers. . London: the Health Foundation, 2009.
111. Michel P. Strengths and weaknesses of available methods for assessing the nature and scale of harm caused by the health system: literature review. 2003.
112. Davis P, Lay-Yee R, Briant R, Schug S, Scott A, Johnson S, et al. Adverse events in New Zealand public hospitals: principal findings from a national study . Ministry of Health, Wellington, New Zealand, 2001.
113. Martins M, Travassos C, Mendes W, Pavao AL. Hospital deaths and adverse events in Brazil. BMC health services research. 2011;11:223.
114. Thomas EJ, Studdert DM, Brennan TA. The reliability of medical record review for estimating adverse event rates. Annals of internal medicine. 2002;136(11):812-6.

Part II

Assessing the quality of medical record review

Chapter 3

The ability of triggers to predict potentially preventable adverse events in a sample of deceased patients

Klein DO, Rennenberg RJMW, Koopmans RP, Prins MH

Preventive Medicine reports 2017;8:250-255

Abstract

Introduction: Several trigger systems have been developed to screen medical records of hospitalized patients for adverse events (AEs). Because it's too labour-intensive to screen the records of all patients, usually a sample is screened. Our sample consists of patients who died during their hospitalization because chances of finding preventable AEs in this subset are highest.

Methods: Records were reviewed for fifteen triggers (n=2182). When a trigger was present, the records were scrutinized by specialized medical doctors who searched for AEs. The positive predictive value (PPV) of the total trigger system and of the individual triggers was calculated. Additional analyses were performed to identify a possible optimisation of the trigger system.

Results: In our sample, the trigger system had an overall PPV for AEs of 47%, 17% for potentially preventable AEs. More triggers present in a record increased the probability of detecting an AE. Adjustments to the trigger system slightly increased the positive predictive value but missed about 10% of the AEs detected with the original system.

Conclusion: In our sample of deceased patients the trigger system has a PPV comparable to other samples. However still, an enormous amount of time and resources are spent on cases without AEs or with non-preventable AEs. Possibly, the performance could be further improved by combining triggers with clinical scores and laboratory results. This could be promising in reducing the costly and labour-intensive work of screening medical records.

Introduction

Unintentional medical harm received increased attention during the past years.[1-10] Several years have passed since the report ''to err is human'' was published, in which the need for a safer health care system was emphasized. Fifteen years after this initial report, a recent update stressed the importance of continuing efforts to improve patient safety.[11] Also, a recent Dutch paper (2013) showed that an average of 12% of patients who died in the hospital still experienced care related injury, which sometimes even contributed to the death of the patient.[12] It is, therefore, important to identify AEs and to determine the risk factors related to their occurrence, in order to reduce harm to patients and improve the quality of care.[13]

It is time-consuming to screen all records for the presence of AEs. Therefore, "triggers" that can be easily identified in the medical records by well-trained nurses in a relatively short time, have been developed. Several trigger systems were created to screen medical records of hospitalized patients for AEs. These triggers are indicators or characteristics of the disease course, known to be often associated with AEs.[14] The fact that cases can be missed, is generally accepted because investigating all records would be too time and cost-consuming in relation to the positive effect of screening. A well-known trigger system is the Global Trigger Tool (GTT), developed by the Institute for Healthcare improvement (IHI). Also, the system from the Harvard medical practice study (HMPS), with a smaller set of triggers, is often used.[15,16] For the aforementioned trigger systems, the positive predictive value has been determined in several studies.[3,17] However, the part of the quality cycle where medical records are scrutinized is still time-consuming and costly. Therefore, it is important to minimize the number of false positive results without increasing the number of false negative results. Because it is too labour-intensive to screen the records of all patients, usually the screening is performed in a sample.

Our sample consists of records of all patients who died during their stay. Therefore, in this study, we used a slightly adapted list of triggers. Examples of cases are illustrated in the supplementary data, to explain some of the most used triggers and the ones which needed extra explanation.

It closely resembles the trigger list from the HMPS, but adjusted to be applicable to medical records of deceased patients. Admittedly, AEs in diseases with negligible mortality but with an unfavourable outcome or hospitalization in departments with low mortality (e.g. ENT, ophthalmology, obstetrics, paediatrics etc.) would escape the opportunity for improvement of care using this sample. Although there are conflicting reports, the most recent and largest study concerning detection of preventable AEs showed that this is particularly effective in deceased patients.[18-20] However, patients who die in hospitals are usually older with more comorbidities and therefore studies in these patients are not generalizable to the average hospital patient. In this study, we assumed that the probability of detecting (serious) AEs was highest in this subset of patients. This would then result in a manageable number of cases to be scrutinized by the committee, but still acquiring a fair overall estimation of the quality of treatment and causes of treatment failure. We wondered whether the positive predictive value (PPV) of the trigger system in deceased patients was acceptable compared to other study samples. Therefore, we analysed our database with informa-

tion on triggers and AEs of all in-hospital deaths in the past years. In addition to this, we performed supplementary analyses in an attempt to optimise the current trigger system.

Methods

This study was performed at the Maastricht University Medical Centre (MUMC+), a teaching hospital in the south of the Netherlands. The medical records used in this study included all inpatient wards including children's. The study protocol was approved by the Ethics Committee of our hospital. We also checked whether patients ever expressed objections against the use of their data for research (this is recorded in a special database in the hospital). If so, their data were excluded. However, none of the patients that were in this sample, did so.

The medical records of all patients who died in our hospital between January 1st, 2012 and January 1st, 2015 were explored by a team of trained nurses for the presence of triggers. Subsequently, a committee consisting of medical specialists from all major disciplines analysed the records to search for AEs. Both the screeners and the specialists were not time restricted. All results were saved using software provided by Medirede®, Clinical File Search version 3 (Mediround BV, 2015). This software was designed to store these data in a clear and easily accessible way. An AE was defined as an unintended outcome arising from the (non)-action of a caregiver and/or the health care system with damage to the patient resulting in temporary or permanent disability or death of the patient.[21] If a potentially preventable AE was suspected, this was discussed with the involved medical department. Finally, the committee decided on the definite presence of an AE and its potential preventability. For the purpose of this study, we used the committee result as a gold standard for AEs. We did not evaluate the effect of hindsight bias, inter- and intrarater reliability.

The starting point of our trigger system was the HPMS list, and we hypothesized that this list would be redundant in deceased patients.[16] Trigger 1 (patient was admitted before (<12 months) for a reason related to the current admission) was adapted to a shorter period (<3 months) because analysis of previous years showed this trigger was not discriminative for potentially preventable AEs. The 12-month cut-off contained a large number of patients with planned chemotherapy or planned second stage operations. Two other triggers were not applicable in a deceased population.

To create a simplified method of triggering, we calculated the positive predictive value (PPV) for the combination of triggers that can be detected by a computer search of the medical records (trigger 1, 4 and 5) and a combination of three triggers that generate the highest number of potentially preventable AEs (trigger 4, 7 and 8). Here, we only looked at the PPV for potentially preventable AEs as the outcome. The PPV of individual triggers was calculated as the rate at which a trigger was associated with an AE, both potentially preventable and not preventable.[22] Furthermore, we calculated risk scores for an AE in patients with a trigger taking the patient characteristics into account. These risk scores could then be used, to generate cut-off points leading to a smaller selection of records with a varying number of AEs depending on the chosen cut-off point.

Statistical analysis

Descriptive statistics are used to describe the general characteristics of the screened medical records and the triggers used in this retrospective analysis.
Chi-square tests and independent t-tests were performed to determine the differences between the groups of patients who experienced an AE during their stay, compared to the group of patients who did not develop an AE.

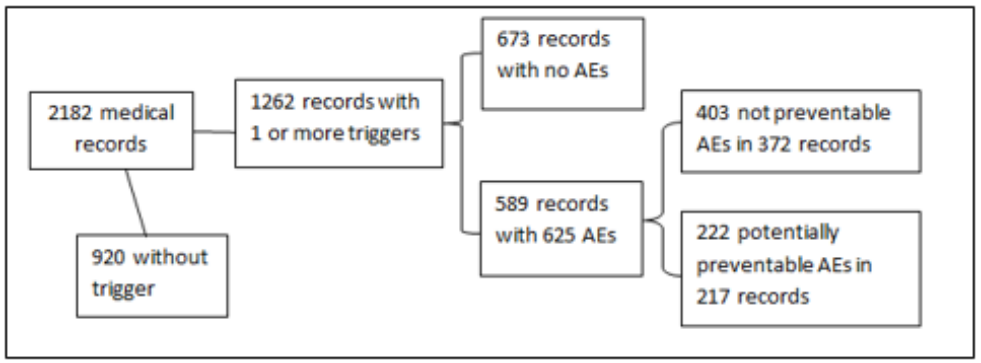
Furthermore, multivariable backward logistic regression analyses (with classification cut-off 0.5) were performed for three scenarios, the first one to detect only computer detectable triggers. The second algorithm contains all 15 triggers to identify the trigger with the highest odds for AEs and potentially preventable AEs. The last algorithm was used to determine the contribution of patient characteristics to the occurrence of AEs to identify possible additional factors that could improve the selection of cases with AEs.

The presence of an AE was used as the dependent variable. Independent variables were: origin (coming from another hospital yes/no), emergency admission, age, gender, admission specialism, and length of stay (in days). Referred by emergency admission was applicable when the patient was admitted via the emergency ward. Admission specialties were divided into surgical (e.g. urology, vascular surgery, gynaecology etc.) and medical departments (e.g. internal medicine, gastroenterology, cardiology, pulmonology, rheumatology, paediatrics etc.). For evaluating the additional value of including the patient characteristics in this last logistic regression algorithm (algorithm 3), we have calculated the probability of every individual of having an AE, given the fact, one or more triggers would be positive. In this algorithm, the following patient characteristics were included: urgent admission, origin, age, gender, length of stay and admission specialism. The logistic regression algorithm yields a continuous outcome, i.e. the predicted probability ranging from 0.0 to 1.0. However, the algorithm will likely be used to classify patients into high risk versus low risk, or positive versus negative. To aid in choosing the right cut-off point for classification, we evaluated 6 different cut-off points. By computing test characteristics for each cut-off point, one cut-off point can be chosen that fits the need for either ruling in or ruling out an adverse event.
Analyses were executed using IBM SPSS Statistics version 23 (IBM Corporation, 2015), a $p<0.05$ was considered statistically significant.

Results

The medical records of 2182 patients were investigated (shown in figure 1). The general characteristics of these patients are shown in table 1. Men were significantly younger than women ($p=0.004$) and they had a significantly higher chance of experiencing an AE ($p=0.021$). The length of stay is significantly longer in patients with an AE compared to patients without an AE ($p<0.001$), whereas preventable and non-preventable AEs don't differ concerning the length of stay ($p=0.911$).

Figure 1 : Flowchart of the medical record analysis in this study



The number of patients admitted to the medical departments is higher than to surgical departments. However, the percentage of patients with an AE is significantly higher at the surgical departments ($p<0.001$).

The PPV of our trigger system is 47%, 589 of the 1,262 positively triggered cases had an AE. 217 of the 1,262 (17%) triggered cases were considered potentially preventable. Table 2 shows the distribution of the individual triggers and their PPVs.

In the triggered cases, the number of unique triggers occurred with a mean of 2.39 per patient (95%CI:2.31-2.46). Finding more triggers gradually increased the likelihood of finding a (potentially preventable) AE. In total 625 AEs were found in 589 records. 33 records showed two or more AEs and 35% of the AEs were potentially preventable. This is shown in table 3.

Table 1: General characteristics of the studied population (patients deceased during hospitalisation)

Variable	Total n=2182 (%)	Trigger present N=1262 (%)	AE present N=589 (%)	Potentially preventable AE present N=217 (%)	Preventable AE potentially contributing to patient death (N=206)
Male	1220 (56)	743 (59)	353 (60)	120 (55)	115
Female	962 (44)	519 (41)	236 (40)	97 (45)	91
Age	69.6 (95%CI 68.8-70.5)	69.2 (95%CI 68.1-70.2)	69.6 (95%CI 68.2-70.9)	71.5 (95%CI 69.3-73.6)	71.5 (95%CI 69.4-73.7)
Length of stay (average days)	12.9 (95%CI 12.2-13.6)	17.1 (95% CI 16.0-18.2)	20.9 (95%CI 19.0-22.9)	21.1 (95%CI 17.6-24.5)	21.3 (95%CI 17.7-24.9)
Urgent admission					
Yes	1546 (71)	1184 (94)	524 (89)	196 (90)	185 (90)
No	636 (29)	78 (6)	65 (11)	21 (10)	21 (10)
Transferred from another hospital					
Yes	186 (9)	104 (8)	47 (8)	19 (9)	18
No	1996 (91)	1158 (92)	542 (92)	198 (91)	188
Admission specialism					
Surgical	455 (21)	351 (28)	257 (44)	107 (49)	106
Medical	1727 (79)	911 (72)	332 (56)	110 (51)	100

*Table 2: Results of individual triggers and AEs**

Triggers		Number of records (% of total)[α]	Non preventable AE	Potentially preventable AE	Total number of AE combined (preventable and non-preventable)	Percentage of total number of AEs	PPV potentially preventable AE (95%CI)~	PPV AE combined (preventable and not preventable) (95%CI)~
1.	Unplanned readmission < 3 months	379 (17.4)	97	47	144	23.0	0.33 (0.25-0.40)	0.38 (0.33-0.43)
2.	Hospital incurred patient injury	116 (5.3)	44	32	76	12.2	0.42 (0.31-0.53)	0.66 (0.57-0.74)
3.	Adverse drug reaction	77 (3.5)	34	13	47	7.5	0.28 (0.14-0.41)	0.61 (0.50-0.72)
4.	Unplanned transfer to ICU	441 (20.2)	158	107	265	42.4	0.40 (0.34-0.46)	0.60 (0.56-0.65)
5.	Unplanned return to the operating room	173 (7.9)	75	67	142	22.7	0.47 (0.39-0.55)	0.82 (0.76-0.88)
6.	Unplanned removal or damage to an organ during surgery.	76 (3.5)	29	29	58	9.3	0.50 (0.37-0.63)	0.76 (0.67-0.86)
7.	Healthcare related infection or sepsis.	509 (23.3)	176	99	275	44.0	0.36 (0.30-0.42)	0.54 (0.50-0.58)
8.	Other complications such as CVA/ lung embolism/acute myocardial infarction/TIA.	350 (16.0)	131	77	208	33.3	0.37 (0.30-0.44)	0.59 (0.54-0.65)
9.	Development of neurological deficit.	129 (5.9)	54	22	76	12.2	0.29 (0.19-0.39)	0.59 (0.50-0.68)
10.	(Initial) unexpected and/or sudden death, absence of terminal care.	266 (12.2)	84	56	140	22.4	0.40 (0.32-0.48)	0.53 (0.47-0.59)
11.	Cardiac or respiratory arrest	197 (9.0)	72	50	122	19.5	0.41 (0.32-0.50)	0.62 (0.55-0.69)
12.	Injury related to abortion or delivery.	-	-	-	-	-	-	-
13.	Dissatisfaction with care.	59 (2.7)	19	11	30	4.8	0.37 (0.18-0.55)	0.51 (0.38-0.64)

14. Documentation indicating litigation.	14 (0.6)	3	6	9	1.4	0.67 (0.28-1.05)	0.64 (0.36-0.93)
15. Other patient complications.	224 (10.3)	57	35	92	14.7	0.38 (0.28-0.48)	0.41 (0.35-0.48)

˜ CI is confidence interval

α Total number of records is 2182.

* The same AE can be found with different triggers, it was not possible to determine which trigger in this case was related to the AE.

ˠ Total number of AEs is 625.

Table 3: Number of triggers and (potentially preventable) AEs with their PPV

Number of triggers (n)	Records (n)	AE present (n)	PPV* for AE (%)	Potentially preventable AE present (n)	PPV* for potentially preventable AE (%)
1	440	134	30.4	45	10,2
2	330	136	41.2	52	15,8
3	224	127	56.7	38	17,0
4	156	106	68.0	42	26,9
5	74	55	74.0	26	35,1
6	26	22	85	9	34,6
7	9	7	78	4	44,4
8	2	1	50	1	50,0
9	1	1	100	0	-

**PPV is positive predictive value*

Detecting AEs with a simple computer algorithm (algorithm 1)

For this analysis, only computer detectable triggers were selected (trigger 1, 4 and 5). 777 cases were positively triggered and contained 391 AEs (this is 63% of the AEs (625) found with the original trigger set). The PPV for an AE with this selection of triggers is therefore 50%. This set of triggers found 147 potentially preventable AEs (66% of potentially preventable AEs (222) found with the original complete trigger set).

When we combine triggers that generate the highest number of potentially preventable AEs (trigger 4, 7 and 8), 162 AEs were found. The PPV for a potentially preventable AE with this system is therefore 20% (this is 73% of potentially preventable AEs (222) found with the original complete trigger set).

Logistic regression with all fifteen triggers (algorithm 2)

In the supplementary data (table 2), the results of individual triggers and AE is shown. The OR was highest for trigger 5 (OR=5.055) and trigger 6 (OR=3.501).

Although not statistically significant, trigger 1 (OR=0.884) and trigger 15 (OR=0.826) suggest a lower risk for finding an AE. For preventable AEs, the OR was the highest for trigger 14 (OR=2.795), trigger 5 (OR=1.671) and trigger 6 (OR=1.588).

Logistic regression with all fifteen triggers and patient characteristics (algorithm 3)

The patient characteristics which were included in the logistic regression combined with the triggers were: urgent admission (yes/no), origin (coming from another hospital yes/no), age, gender, the length of stay (days) and admission specialism (surgical/medical).
In the supplementary data (table 4 and 5), the results of the combination of all triggers and these additional characteristics are shown.
To find out which combination could identify the highest number of patients having an AE,

we chose several cut-off points, which are shown in table 4. A cut-off point of 0.3 would mean that 532 of the 589 cases with AE would be found and that fewer records were selected for review than in the original review (PPV is therefore 50%). With a cut-off point of 0.3, 194 of the 217 possibly preventable AEs will be detected. Higher cut-off points detected less AEs.

Table 4: Cut-off points

Cut-off point	Number of medical records selected	AE found	Cases with AE missed	Potentially preventable AE found	Cases with potentially preventable AE missed
≥0.1	1262	589	0	217	0
≥0.2	1259	589	0	217	0
≥0.3	1061	532	57	194	23
≥0.4	537	360	229	134	83
≥0.5	437	310	279	118	99
≥0.6	386	281	308	110	107

Discussion

This study showed that the trigger system had an average PPV for AEs of 47% and for potentially preventable AEs of 17%. The more triggers found in a case, the higher the probability of finding an AE. Adjustments to the trigger system slightly increased the PPV for AEs and potentially preventable AEs but fail to identify around 10% of cases (cut-off point 0.3) compared to the complete original trigger system.
The triggers of the ''Harvard Medical Practice Study'' and the "Global Trigger Tool'' have an overall PPV of 40.3% and 30.4% respectively.[17] This matches well with the results in our sample. Looking at the individual triggers we found that unplanned removal, damage or repair of an organ (PPV=76.3% for total AE) and unplanned return to the operating room (PPV=82.1% for total AE) had the highest predictive value for an AE. A study by Naessens et al (2010) also showed that the trigger with the highest yield was 'return to the operating room' where 80.6% of these patients suffered from an AE.[22]. Hwang et al (2014) analysed the global trigger tool. They found that only six triggers had positive predictive values of greater than 50%.[23] Two of these PPV's could be reproduced by the triggers in our data. The definitions of the other triggers were not comparable. 12 of our triggers had a PPV higher than 50%. Possibly this is caused by a difference in patient selection (we only investigated deceased patients) or in the expertise of the committee that investigated the records and adjudged AEs.

Clinical and patient characteristics associated with increased occurrence of AEs were admission through the emergency room, transfer from another hospital, a higher number of triggers and admission for a surgical specialism. Although the first three seem logical the latter suggests a higher risk in surgical wards that has no easy explanation. This has been noticed in other studies.[4] Freund et al (2013) also found no significant difference in terms of age and sex.[24] However, the data were not corrected for comorbidities or the condition

of the patients. Furthermore, the complications are usually closely related to a surgical intervention making the allocation of an AE to the intervention easy, whereas in medical specialties the complications of, for instance, pharmacological interventions, are less predictable or occur later making the detection of a link more difficult. Moreover, surgeons are ahead with registration of complications compared to medical departments and this might simplify the finding an AE in these patients.[25,26]

One would expect that cases with multiple triggers or a longer duration of stay in the hospital experience more AEs. Our study indeed shows a higher risk for AEs as the number of triggers per case increases. Although there seems to be a trend in more risk for an AE with longer hospitalization, this was not statistically significant for preventable AEs.

A report published by NIVEL (The Netherlands) in 2013 showed that on average 12% of all patients who die in the hospital experience an AE. In academic hospitals, this was 15.1% (95%CI 11.8-19.0). Four percent of the patients who died during their stay experienced a preventable AE according to this report (2.3% in academic hospitals).[27]

Our study found that 27% of all patients who died in the hospital experienced an AE of which 37% (10% of all patients) was considered potentially preventable. This might partially be explained by the fact that we, in contrast to NIVEL, also incorporated gynaecology, psychiatry and neonatology cases. Although the subjects in the NIVEL study were older (75 vs. 69.6 years) they had a shorter length of stay (10.4 vs. 12.9 days) indicating this sample might not be completely comparable to ours. Furthermore, internal reviewers are known to find more AEs than external reviewers and NIVEL only recruited external reviewers which had a medical, surgical, or neurological background.[28,29] A certain level of bias in judging single cases cannot be excluded and might also increase the number of AEs found. Internationally preventability of AEs ranges from 43 to 70%.[30-32] This indicates that there are considerable differences in the way judgments about preventability, or even the presence of an AE, are made. Although some studies use different grades of preventability, there is no international consensus on how to specifically apply these grades. Therefore, comparisons can hardly be made in view of different methods used.

Combining triggers and clinical characteristics seems promising in reducing the review of cases without an AE. However, some (potentially preventable) AEs will be missed. Quality and safety departments in hospitals have to decide on the optimal cut-off point. This could then result in less medical records necessary to be screened by the specialist, saving a fair amount of time and costs.

Possibly entering additional variables into the system might increase the gain of this system. For example, clinical scores of vital functions (like modified early warning score) or laboratory results (like albumin, creatinine, haemoglobin level etc.). These might be combined with the existing triggers to improve the PPV.

The strength of our study is the specific and reliable recording of triggers and AEs using software specifically developed for this purpose. Also, the number of cases from a single hospital over a period of 3 years is large enough to generate reliable results with small confidence intervals. Furthermore, there was a stable trigger team during the period se-

lected for this study and the presence of an AE was decided on after discussion within the committee generating broad support for the final decision. Preventability was judged by the consensus of several doctors with a variable background. In the face of a lacking international consensus on this concept of preventability, this seems an optimal and acceptable method.

Clearly, there are also points for improvement. We compared formerly reported PPV's of these two commonly used trigger sets in different patient samples with the result of a slightly adapted HPMS trigger set in our sample of deceased patients. We have a single measurement of our PPV over a period of three years whereas the PPV of the other trigger systems is based on an average of several studies presented in the literature. Furthermore, we have no information on the negative predictive value of our trigger system. It is possible that some of the cases that were not triggered contain AEs or potentially preventable AEs.

We realise that we looked into a subset of patients which makes our results not generalizable to the average hospital patient. He or she might be younger with fewer comorbidities and other diseases in different departments. Therefore, important potentially preventable AEs in these patients could have been missed. Our data also gave us no information about the reproducibility of the trigger system. For the HMPS method, Kappa values are reported between 0.5333 and 0.7634 (moderate to good agreement), for the IHI method between 0.2035 and 0.7836 (slight to good agreement). Lastly, it is likely that in a confirmation study that the PPVs will be lower than we have found in this derivation study.

In our opinion, it is disappointing that trigger systems select over 50% of cases without an AE. Even after combining several triggers the PPV does not significantly improve. This method remains, therefore, labour-intensive until we can define triggers or trigger sets with a higher PPV. Further research to optimise these systems concerning the combination of triggers with patient characteristics or possible even laboratory results seems warranted. Due to the expected higher number of AEs in deceased patients, we expected this tool to perform better in this subsample of patients. However, we think that the PPV of the HPMS in this sample is disappointing but compares well to results from general inpatient samples using the HPMS or IHI trigger system.

References

1. Rutberg H, Borgstedt Risberg M, Sjodahl R, Nordqvist P, Valter L, Nilsson L. Characterisations of adverse events detected in a university hospital: a 4-year study using the Global Trigger Tool method. BMJ open. 2014;4(5):e004879.
2. Najjar S, Hamdan M, Euwema MC, Vleugels A, Sermeus W, Massoud R, et al. The Global Trigger Tool shows that one out of seven patients suffers harm in Palestinian hospitals: challenges for launching a strategic safety plan. Int J Qual Health Care. 2013;25(6):640-7.
3. Kennerly DA, Saldana M, Kudyakov R, da Graca B, Nicewander D, Compton J. Description and evaluation of adaptations to the global trigger tool to enhance value to adverse event reduction efforts. J Patient Saf. 2013;9(2):87-95.
4. Zegers M, de Bruijne MC, Wagner C, Hoonhout LH, Waaijman R, Smits M, et al. Adverse events and potentially preventable deaths in Dutch hospitals: results of a retrospective patient record review study. Qual Saf Health Care. 2009;18(4):297-302.
5. Mull HJ, Brennan CW, Folkes T, Hermos J, Chan J, Rosen AK, et al. Identifying Previously Undetected Harm: Piloting the Institute for Healthcare Improvement's Global Trigger Tool in the Veterans Health Administration. Quality management in health care. 2015;24(3):140-6.
6. Kurutkan MN, Usta E, Orhan F, Simsekler MC. Application of theIHI Global Trigger Tool in measuring the adverse event rate in a Turkish healthcare setting. Int J Risk Saf Med. 2015;27(1):11-21.
7. Farup PG. Are measurements of patient safety culture and adverse events valid and reliable? Results from a cross sectional study. BMC health services research. 2015;15:186.
8. Welfare NIfHa. IHI Global Trigger Tool and patient safety monitoring in Finnish hospitals - Current experiences and future trends. 2013.
9. Goodman JC, Villarreal P, Jones B. The social cost of adverse medical events, and what we can do about it. Health affairs (Project Hope). 2011;30(4):590-5.
10. Baines RJ, Langelaan M, de Bruijne MC, Wagner C. Is researching adverse events in hospital deaths a good way to describe patient safety in hospitals: a retrospective patient record review study. BMJ Open. 2015;5(7):e007380.
11. Foundation NPS. Free from Harm: Accelerating Patient Safety Improvement Fifteen Years after To Err Is Human 2015.
12. Langelaan MdBM, C; Baines, R.J; Broekens, M.A; . Monitor zorggerelateerd schade 2011/2012. Dossieronderzoek in Nederlandse ziekenhuizen. . 2013.
13. Hwang JI, Chin HJ, Chang YS. Characteristics associated with the occurrence of adverse events: A retrospective medical record review using the Global Trigger Tool in a fully digitalized tertiary teaching hospital in Korea. Journal of Evaluation in Clinical Practice. 2014;20(1):27-35.
14. Resar RK, Rozich JD, Classen D. Methodology and rationale for the measurement of harm with trigger tools. Qual Saf Health Care. 2003;12 Suppl 2:ii39-45.
15. Griffin FA RRC, MA: . IHI Global Trigger Tool for Measuring Adverse Events (Second Edition) IHI Innovation Series white paper. Cambridge, Massachusetts: Institute for Healthcare Improvement. 2009.
16. Brennan TA, Leape LL. Adverse events, negligence in hospitalized patients: results from the Harvard Medical Practice Study. Perspect Healthc Risk Manage. 1991;11(2):2-8.
17. Unbeck M, Schildmeijer K, Henriksson P, Jurgensen U, Muren O, Nilsson L, et al. Is detection of adverse events affected by record review methodology? an evaluation of the "Harvard Medical Practice Study" method and the "Global Trigger Tool". Patient Saf Surg. 2013;7(1):10.
18. Baines RJ, Langelaan M, de Bruijne MC, Wagner C. Is researching adverse events in hospital deaths a good way to describe patient safety in hospitals: a retrospective patient record review study. Bmj Open. 2015;5(7).
19. Hogan H, Healey F, Neale G, Thomson R, Vincent C, Black N. Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study. BMJ Qual Saf. 2012;21(9):737-45.
20. Dunn KL, Reddy P, Moulden A, Bowes G. Medical record review of deaths, unexpected intensive care unit admissions, and clinician referrals: detection of adverse events and insight into the system. Arch Dis Child. 2006;91(2):169-72.
21. Wagner C. Onbedoelde schade in ziekenhuizen: resultaten dossieronderzoek naar patiëntveiligheid. Klachtenmanagement in de Zorg. 2007;4(3-4):28-31.
22. Naessens JM, O'Byrne TJ, Johnson MG, Vansuch MB, McGlone CM, Huddleston JM. Measuring hospital adverse events: Assessing inter-rater reliability and trigger performance of the Global Trigger Tool. International Journal for Quality in Health Care. 2010;22(4):266-74.
23. Hwang JI, Chin HJ, Chang YS. Characteristics associated with the occurrence of adverse events: a retrospective medical record review using the Global Trigger Tool in a fully digitalized tertiary teaching hospital in Korea. J Eval Clin Pract. 2014;20(1):27-35.
24. Freund Y, Goulet H, Bokobza J, Ghanem A, Carreira S, Madec D, et al. Factors associated with adverse

events resulting from medical errors in the emergency department: two work better than one. J Emerg Med. 2013;45(2):157-62.
25. Marang-van de Mheen PJ, Kievit J. [Automated registration of adverse events in surgical patients in the Netherlands: the current status]. Ned Tijdschr Geneeskd. 2003;147(26):1273-7.
26. Beleidsdocument Complicatieregistratie Nederlande Internisten Vereniging, 2010.
27. Langelaan M, Bruijne de, M. C., Baines, R.J., . Monitor zorggerelateerde schade 2011/2012 - Dossiersonderzoek in nederlandse ziekenhuizen 2013.
28. Langelaan M. Monitor zorggerelateerde schade 2011/2012 - Dossieronderzoek in Nederlandse ziekenhuizen. EMGO+ Instituut/VUmc, NIVEL, Nederlands Instituut voor onderzoek van de gezondheidszorg, 2013.
29. Landrigan CP, Parry GJ, Bones CB, Hackbarth AD, Goldmann DA, Sharek PJ. Temporal trends in rates of patient harm resulting from medical care. N Engl J Med. 2010;363(22):2124-34.
30. Wang CH, Shih CL, Chen WJ, Hung SH, Jhang WJ, Chuang LJ, et al. Epidemiology of medical adverse events: perspectives from a single institute in Taiwan. Journal of the Formosan Medical Association = Taiwan yi zhi. 2016.
31. Aranaz-Andres JM, Aibar-Remon C, Vitaller-Murillo J, Ruiz-Lopez P, Limon-Ramirez R, Terol-Garcia E, et al. Incidence of adverse events related to health care in Spain: results of the Spanish National Study of Adverse Events. J Epidemiol Community Health. 2008;62(12):1022-9.
32. von Laue NC, Schwappach DL, Koeck CM. The epidemiology of medical errors: a review of the literature. Wiener klinische Wochenschrift. 2003;115(10):318-25.
33. Soop M, Fryksmark U, Koster M, Haglund B. The incidence of adverse events in Swedish hospitals: A retrospective medical record review study. International Journal for Quality in Health Care. 2009;21(4):285-91.
34. Wilson RM, Michel P, Olsen S, Gibberd RW, Vincent C, El-Assady R, et al. Patient safety in developing countries: retrospective estimation of scale and nature of harm to patients in hospital. BMJ. 2012;344:e832.
35. Schildmeijer K, Nilsson L, Arestedt K, Perk J. Assessment of adverse events in medical care: lack of consistency between experienced teams using the global trigger tool. BMJ Qual Saf. 2012;21(4):307-14.
36. O'Leary KJ, Devisetty VK, Patel AR, Malkenson D, Sama P, Thompson WK, et al. Comparison of traditional trigger tool to data warehouse based screening for identifying hospital adverse events. BMJ quality & safety. 2013;22(2):130-8

Supplementary data

I: Examples cases

Examples cases are illustrated, to explain some of the most used triggers and the ones which needed some extra explanation.

Table 1: A list of the triggers which are used in the MUMC+

Unplanned readmission after discharge from index admission within 3 months
Hospital-incurred patient injury (temporarily or lasting)
Adverse drug reaction
Unplanned transfer to ICU
Unplanned return to the operating room
Unplanned removal, injury or repair of an organ during surgery
Healthcare related infection or sepsis
Other complications such as CVA/pulmonary embolism
Development of neurological deficit not present on admission
(initial) unexpected and/or sudden death (not palliative care)
Cardiac or respiratory arrest
Injury related to abortion or delivery
Dissatisfaction with care
Documentation indicating litigation
Other patient complications

Case A:

A 50-year-old male was admitted for the second time within 4 weeks because of decompensated alcoholic liver disease. His previous history reports epilepsy after excision of a meningioma. His anti-epileptic drugs were recently changed because of his liver failure. During hospitalization he suffered a seizure with aspiration making transfer to the ICU necessary. Despite maximal treatment including intubation there was no improvement and the patient died several days later.

- Trigger 1 re-admittance
- Trigger 4 unexpected transfer to the ICU
- Trigger 15 seizure possibly related to medication change?

Case B:

A 78-year-old male was admitted because of chest-pain. Two vessel coronary disease and an aortic stenosis with a gradient of 40 mmHg were diagnosed. After CABG with aortic valve replacement the patient remained unstable on the ICU due to cardiac tamponade. There was a re-operation were blood was removed from the pericardium but no clear bleeding focus was identified. Thereafter, his condition deteriorated quickly with intestinal ischemia for which a colectomy was done. Despite this and maximal treatment he died one day later.

- Trigger 5 unplanned return to the operating room

Case C:
A 89-year-old male was admitted because of cholecystitis due to an obstructing gallstone in the neck of the gallbladder. He was treated with an external drain and amoxicillin/clavulanic acid. His condition improved gradually. After 4 days he was found in bed early in the morning with a Glasgow coma score of 1-1-1. A cerebral CT-scan revealed a large subdural hematoma with midline shift. He died the same day.

- Trigger 10 Unexpected death

Case D:
A 93-year-old male was admitted because of sudden delirium. His previous history revealed atrial fibrillation for which he used warfarin. The primary analysis, including CT of the brain, did not reveal a direct cause. On day four he fell to the floor on his way to the toilet. He hit his head and a new CT scan showed a subdural hematoma. His INR was 5.5 and he immediately received coagulation factor concentrate. Despite full correction of the INR he deteriorated quickly. A neurosurgical intervention was against his and his family's wishes and he died the same day.

- Trigger 2 Hospital incurred injury

II: SPSS output

Table 2: SPSS results for logistic regression of all triggers in relation to AE (n=1262)*

							95% C.I.for EXP(B)	
	B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Trigger 1	-,123	,142	,752	1	,386	,884	,670	1,168
Trigger 2	,700	,225	9,680	1	,002	2,014	1,296	3,131
Trigger 3	,488	,261	3,483	1	,062	1,628	,976	2,717
Trigger 4	,592	,134	19,450	1	,000	1,807	1,389	2,351
Trigger 5	1,620	,219	54,927	1	,000	5,055	3,293	7,759
Trigger 6	1,253	,298	17,643	1	,000	3,501	1,951	6,283
Trigger 7	,220	,135	2,665	1	,103	1,246	,957	1,622
Trigger 8	,600	,144	17,448	1	,000	1,822	1,375	2,414
Trigger 9	,279	,210	1,762	1	,184	1,322	,876	1,994
Trigger 10	,615	,157	15,335	1	,000	1,850	1,360	2,517
Trigger 11	,469	,179	6,889	1	,009	1,598	1,126	2,268
Trigger 13	,141	,295	,229	1	,632	1,151	,646	2,051
Trigger 14	,159	,652	,059	1	,808	1,172	,327	4,206
Trigger 15	-,191	,165	1,331	1	,249	,826	,598	1,142
Constant	-1,121	,134	70,109	1	,000	,326		

*Definition of the triggers is described in table 1.

Table 3: SPSS results for logistic regression of all triggers in combination with patient characteristics in relation to potential AE (n=1262)*

							95% CI for EXP(B)	
	B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Urgent admission	-1,204	,339	12,573	1	,000	,300	,154	,584
Origin	-,724	,249	8,467	1	,004	,485	,298	,789
Age (years)	,002	,004	,318	1	,573	1,002	,995	1,009
Gender	,003	,133	,000	1	,985	1,003	,773	1,301
Length of stay	,012	,004	8,935	1	,003	1,012	1,004	1,020
Admission specialism	1,205	,157	58,777	1	,000	3,336	2,452	4,539
Trigger 1	-,054	,149	,131	1	,717	,947	,708	1,269
Trigger 2	,625	,238	6,907	1	,009	1,868	1,172	2,976
Trigger 3	,517	,275	3,546	1	,060	1,678	,979	2,875
Trigger 4	,534	,141	14,266	1	,000	1,706	1,293	2,251
Trigger 5	1,250	,233	28,801	1	,000	3,490	2,211	5,510
Trigger 6	1,129	,315	12,871	1	,000	3,092	1,669	5,728
Trigger 7	,108	,153	,500	1	,480	1,114	,825	1,505
Trigger 8	,489	,151	10,545	1	,001	1,631	1,214	2,192
Trigger 9	,377	,220	2,937	1	,087	1,459	,947	2,246
Trigger 10	,592	,165	12,872	1	,000	1,808	1,308	2,500
Trigger 11	,477	,192	6,204	1	,013	1,612	1,107	2,347
Trigger 13	,145	,309	,221	1	,638	1,157	,631	2,121
Trigger 14	,086	,714	,014	1	,905	1,089	,269	4,415
Trigger 15	-,265	,174	2,322	1	,128	,767	,546	1,079
Constant	-,445	,456	,952	1	,329	,641		

*Definition of the triggers is described in table 1.

Table 4: SPSS results for logistic regression of all triggers in combination with patient characteristics in relation to potential preventable AE (n=589)*

							95% CI for EXP(B)	
	B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Urgent admission	,487	,308	2,495	1	,114	1,627	,889	2,978
Origin	,220	,340	,417	1	,518	1,246	,640	2,427
Age (years)	,013	,006	4,121	1	,042	1,013	1,000	1,025
Gender	-,290	,183	2,512	1	,113	,748	,523	1,071
Length of stay	,000	,004	,012	1	,914	1,000	,991	1,008
Admission specialism	,162	,193	,704	1	,401	1,176	,805	1,718
Trigger 1	-,136	,217	,392	1	,531	,873	,571	1,335
Trigger 2	,245	,268	,836	1	,361	1,278	,755	2,161
Trigger 3	-,373	,353	1,116	1	,291	,689	,345	1,375
Trigger 4	,296	,186	2,535	1	,111	1,345	,934	1,936
Trigger 5	,504	,219	5,288	1	,021	1,656	1,077	2,545
Trigger 6	,431	,299	2,078	1	,149	1,539	,856	2,766
Trigger 7	-,061	,212	,082	1	,774	,941	,621	1,426
Trigger 8	,022	,192	,013	1	,910	1,022	,702	1,488
Trigger 9	-,352	,282	1,557	1	,212	,703	,404	1,223
Trigger 10	,205	,215	,905	1	,341	1,227	,805	1,872
Trigger 11	,187	,224	,695	1	,405	1,205	,777	1,870
Trigger 13	-,186	,417	,199	1	,656	,830	,367	1,880
Trigger 14	1,195	,754	2,512	1	,113	3,304	,754	14,482
Trigger 15	,087	,249	,121	1	,728	1,091	,669	1,778
Constant	-2,096	,597	12,336	1	,000	,123		

*Definition of the triggers is described in table 1.

Chapter 4

The Harvard Medical Practice Study trigger system performance in deceased patients

Klein DO, Rennenberg RJMW, Koopmans RP, Prins MH

BMC Health Services Research Jan 8;19(1):16.2019

Abstract

Introduction: To detect possible threats to quality and safety, multiple systems have been developed. One of them is retrospective chart review. A team of experts scrutinizes medical records, selected by trigger systems, to detect possible adverse events (AEs). The most important AEs and more hints for possible improvement of care appear in deceased patients. Using triggers in a sample of these patients might increase the performance and lower the burden of scrutinizing records without possible preventable AEs. The aim of this study was therefore to determine the performance of the trigger system in a sample of deceased patients and to calculate the specificity and the sensitivity of this trigger system for predicting AEs.

Methods: We performed a study in which the records of deceased patients were screened for triggers by a team of trained nurses. A sample of 100 medical records was randomly selected out of records which had been screened before, prior to the study in 2016. For the determination of significant differences between the first and second screening, McNemar's test of symmetry was used. Also, observed agreement, Cohen's Kappa and prevalence-adjusted and-bias-adjusted-kappa (PABAK) statistics were calculated. This was done for the two trigger rounds on both any trigger present and for every trigger separately.

Results: The observed agreement for any given trigger was 75% with a Kappa and PABAK of 0.5. For the individual triggers, the observed agreement was on average 90%. The corresponding Kappa was on average 0.42 (range: -0.03-0.78) and the average PABAK was 0.8 (range: 0.44-0.92). Two adverse events were found in cases without triggers previously. The recalculated specificity and sensitivity for the original population were 58% and 92% respectively.

Conclusion: For the reproducibility of triggers it seems that some perform better than others, but on average this is to our opinion suboptimal. The low specificity implies that many records are selected without AEs. This leads to a high false-positive rate making this labour-intensive record review process costly. Therefore, research for better and more expedient systems is required.

Introduction

Improving quality and safety of care in hospitals has become an important focus of health care policy in the past decades. This was initiated by reports such as ''to err is Human'' (1999), and in the Netherlands by the report "adverse events in Dutch hospitals" (2004).[1,2] The latter study was repeated in 2008 and 2012. Although there was an improvement, still a considerable number of (potentially preventable) adverse events (AEs) was found.[3,4] Also a report by Landrigan et al (2010) stated that further efforts are necessary to improve safety strategies and to monitor health care safety over time.[5] To detect possible threats to quality and safety, multiple systems have been developed. One of them is retrospective chart review. A team of experts scrutinizes medical records to detect possible AEs. The involved departments should then be able to learn from these events and improve their care by increasing awareness and adapting protocols or guidelines.

It is clear that screening of the medical records of all patients by specialists is time-consuming and costly.[6] Therefore, trigger systems have been developed to select cases in which an AE is probably present.[7] Triggers are clues which alert screeners for potential AEs (for example "unplanned transfer to the intensive care unit''). The medical record can then be thoroughly reviewed to determine if an actual AE has occurred. There are two main trigger systems used widely and the triggers are usually applied to the medical files by trained screeners in both systems. The first one was developed for the Harvard Medical Practice Study (HMPS) study and has 18 triggers.[8] Thereafter, the Institute for Healthcare Improvement (IHI) tried to improve the performance of this trigger tool and developed their system with 54 triggers.[9] Both systems are used for retrospective medical record review. However, in contrast to the IHI trigger system, the HMPS method is mostly used for research purposes.[10] Although the HMPS trigger set is rather old, it is still used in national screening programs to evaluate patient safety in hospitals.[10-13]

Usually, these trigger systems are applied to records of discharged patients which are functioning well. However, the literature shows that the most important AEs and more hints for possible improvement of care in all patients appear in deceased patients.[14-17] Moreover, several studies have shown higher numbers of preventability of AEs in this subgroup.[2,12,14,18] Thus, apart from optimizing the trigger tool itself, using it in this sample of patients might increase the performance and lower the burden of scrutinizing records without possible preventable AEs.[19] Finding AEs and possible points of improvement is considered important by our hospital and therefore medical record review of deceased patients has been applied for many years.

We wondered if the performance of the trigger system would be better in a sample of deceased patients because it was shown that in these patients more triggers can be easily found.[20] Therefore, we performed a study in which the records of deceased patients were screened twice for these triggers by the same team of trained nurses. Our aim was to determine the agreement for the two trigger rounds on both any trigger present and for every trigger separately using the HMPS trigger system in deceased patients.

Methods

Medical record review

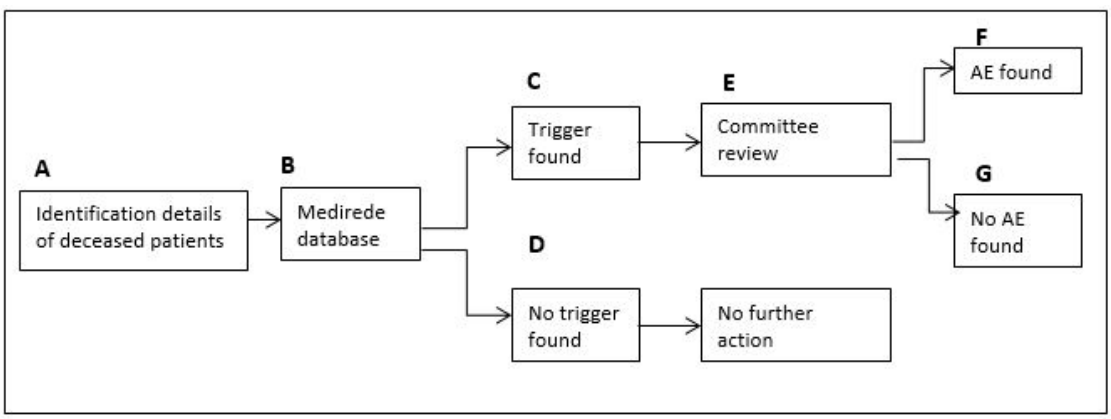
Screeners

Since 2008 a stable team of seven nurses with broad clinical experience (all working more than 10 years in their clinical department) has screened the medical records of all deceased patients for the presence of triggers from the HMPS system. They were trained in record review and patient safety initially for the national monitoring on triggers and AEs as described by the Netherlands Institute for Health Services Research (NIVEL).[21] During their work at the intensive care and emergency departments, they have been confronted frequently with dying patients. This has resulted in them being pre-eminently suitable to judge events concerning these patients.

Trigger system

Patient identification details of those who died during their stay were uploaded in a secured software program designed to aid medical record review of deceased patients (Medirede®, Clinical File Search version 3; Mediround BV, 2015). Then the matching records were screened by one of the nurses from the team using the HPMS trigger system. If a medical record contained at least one trigger, it was regarded as positive and was then forwarded to the review committee and scrutinized for the potential presence of AEs. This is shown in figure 1.

Figure 1: Procedure of medical record review in our centre



(A) Identification details of all deceased patients are inserted in the Medirede® database (B) and the corresponding medical record is screened by one of the nurses. (C) When one or more triggers are found in the medical record (registered in Medirede), the case is flagged and will continue to the review committee.
(D) When no trigger is found (also registered in Medirede), no further action will be taken in the normal screening procedure. (E) The committee evaluates the cases and will determine whether (F) an adverse event occurred or (G) not.

We used a slightly adapted version of the HMPS trigger system to make it suitable for the screening of deceased patients.[22] The triggers regarding transfer to another acute care hospital and unplanned inappropriate discharge to home were omitted as they have no relevance in deceased patients. The same triggers were used throughout the whole period except the trigger regarding readmission of the patient which was changed in 2013 (originally; the patient was admitted before (<12 months) for a reason related to index admission). Analysis of our database showed that this trigger was not discriminative for (potentially preventable) AEs (data not shown). For example: within the 12-month period, many patients were selected with planned repeated chemotherapy or planned reversal operations (which by their very nature are not related to AEs and thus not useful for our purposes). Therefore, the definition of trigger 1 was adapted to ''patient has been admitted in the previous three months for a reason related to the index admission''. Example of cases with corresponding triggers is explained in further detail elsewhere.[23] Each record was reviewed twice by one of the nurses, once prior to the study in the context of their regular work as screeners and once during this study.

Review committee

Usually, only medical records with a trigger were forwarded to one of the members of the review committee. The committee consisted of several clinical specialists representing the departments with most of the in-hospital deaths who were trained to identify AEs according to NIVEL standards.[21] For this study, they also evaluated records without a trigger (figure 1, D). After an evaluation of the medical record, they decided together (in a consensus meeting) whether an AE had occurred during the hospitalisation of the patient.

Study

This study was performed in 2016 at the Maastricht University Medical Centre (MUMC+), a large teaching hospital in the south of the Netherlands. The medical records that were used in this study included a sample of all inpatient deaths between January 2012 and January 2015 (in total 2182 cases). The study protocol was approved by the medical ethics committee of our hospital.

Data and analysis

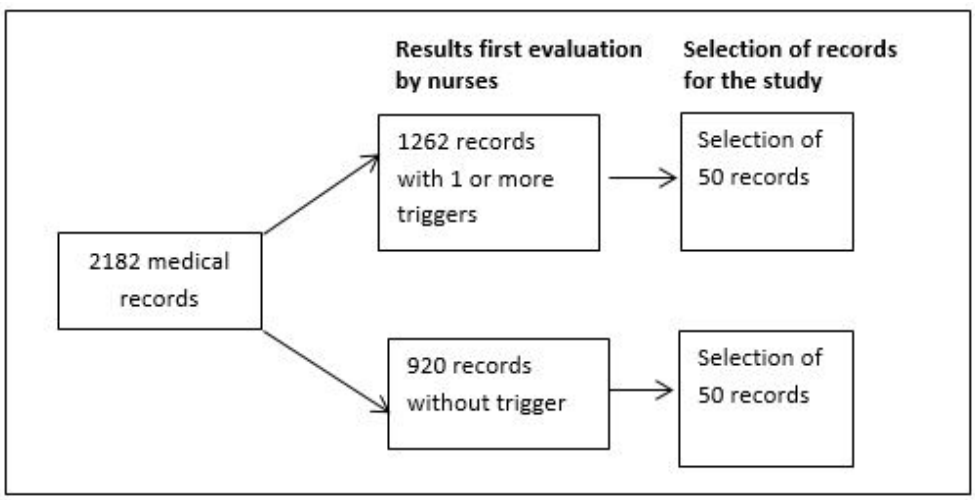
We aimed to get a point estimate with an exact 95% confidence level and a confidence interval of 5% to each side, based on the total number of records. Hence, we needed to include at least 92 cases. Therefore, a sample of 100 medical records was randomly selected (with the use of Excel's random generator) out of records which had been screened in the preceding years. Characteristics of the patient sample are presented in table 1.

Table 1: General characteristics of the patient sample (2012-2015)

Average age (years)	67.6
Gender	52% male
Admission specialism	5% paediatrics
	7% other
	9% neurology
	12% lung diseases
	14% ICU
	14% cardiology
	15% surgery
	24% internal medicine
Average length of stay (days)	13.5

We selected fifty of these records from the set without triggers in the first screening and fifty from the set with at least one trigger present. The study flow is depicted in figure 2. To ensure that the nurses were blinded to the results of the first screening we changed the ID numbers of the records, making it impossible to consult previous results.

Figure 2: Study flow



In the primary analysis, the nurses were analysed as a group instead of as individuals. If a small subsample of at least ten cases was triggered by the same nurse during the first and second round, we calculated the Kappa also separately in this subgroup analysis.
For the determination of significant differences between the two screening rounds, McNemar's test of symmetry was performed. Observed agreement and Cohen's Kappa statistics (with 95% confidence interval; CI) were calculated between the two trigger rounds on both any trigger present and for every trigger separately. For the calculation of the observed

agreement (reliability), we divided the total number of cases with a comparable judgment in both screening rounds by the total number of reanalysed records (100). We also checked whether there was a difference between objective triggers (trigger 1, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 14) and subjective triggers (trigger 2, 12, 13, 15).
Furthermore, we calculated prevalence-adjusted and bias-adjusted kappa (PABAK) and reported this along with Kappa, to show how data would have been with equal distributions of positive and negative test results. Finally, we determined prevalence-indexes and bias-indexes.[24]

Analyses were carried out using IBM SPSS Statistics version 23 (IBM Corporation, 2015). A $p<0.05$ was indicated as significant.

All selected records were evaluated by the review committee (regardless if triggers were found by our screeners). We recalculated the numbers of triggers and AEs to represent the original number of patients in the specific population. With this information, we were able to calculate the specificity and sensitivity of the trigger system.

Results

Results for any trigger present

Table 2 shows that the second screening revealed 20 new cases with a trigger. 45 cases had a trigger in both screening rounds. This resulted in 65 records with a trigger in this study. After the screening in the study, 35 records didn't have a trigger. 30 of these records didn't have a trigger in both screening rounds. 5 of these records had a trigger after the first screening but remained without trigger after the screening in the study.

The observed agreement for the triggers was 75%, and the corresponding Cohen's Kappa was 0.50 (95%CI:0.34-0.66). PABAK was calculated as 0.5 (95%CI:0.29-0.66). An exact McNemar's test confirmed that there was a significant difference in the proportion of positively triggered records the first and second time, with p=0.004.

Table 2: Numbers of triggers and AEs in the first and second round

Trigger first round	Trigger second round (study)	Number of cases	AE first round	AE second round (study)
+	-	5	0	0
+	+	45	25	28
-	+	20	NA[1]	5
-	-	30	NA[1]	2

[1]NA: not assessed; because there was no trigger in the first round, these records were not investigated in the first round.

AEs found in relation to trigger status

In the 20 cases with a newly detected trigger, 5 AEs were found. In the cases without a trigger in both rounds 2 AEs, were found. After recalculating the numbers of these proportions to represent the whole population we found a specificity of 58% (95% CI 55.7-60.8) and a sensitivity of 92% (95% CI 90.1-94.2) for detecting AEs.

Results for the individual triggers

In table 3, the results are shown for the individual triggers during the first and second screening. Trigger 12 was not present in both screening rounds. Eleven triggers were more often detected during the assessment in this study. Kappa for agreement ranged between 0.03 and 0.78 for the individual triggers, with an average of 0.42. The average PABAK was 0.80. The observed agreement was on average 90%.

McNemar's test was executed for every single trigger but was only significant for trigger 15 ($p<0.0001$). Trigger 15 was significantly more found in the second round. Furthermore, the objective triggers had a higher average Kappa (0.46) compared to the subjective triggers (average K= 0.28). The corresponding PABAK was 0.82 and 0.71, respectively.

Subanalyses

These subanalyses were executed because twelve cases were analysed by the same nurse during the two screenings. For any trigger present, the Kappa was 0.63 (95%CI:0.15-1), PABAK was 0.67 (95%CI:0.25-1). The average Kappa of the individual triggers was 0.70 and the average PABAK was 0.83.

If these twelve cases would be excluded from the total analyses, the Kappa for any trigger present would be 0.48 (95%CI:0.30-0.67) (compared to 0.5 with these 12 included) and PABAK would be 0.48 (95%CI:0.29-0.66). The average Kappa for the individual triggers would be 0.39 (compared to 0.42 in the total sample) and corresponding PABAK 0.78.

Table 3: Triggers given the first and second time of screening

Triggers	Description	First time positive (N)[1]	Second time positive (N)[1]	Percentage agreement (%)	Kappa agreement (95%CI)	McNemar's test (p-value)	PABAK[2] (95%CI)	Bias index	Prevalence index
Trigger 1	Unplanned readmission[3] (within 3 months) after discharge from index admission	13	17	88	0.53 (0.30-0.76)	0.39	0.76 (0.58-0.88)	-0.04	-0.7
Trigger 2	Hospital-incurred patient injury (temporarily or lasting)	8	4	90	0.12 (-0.17-0.41)	0.34	0.80 (0.63-0.90)	0.04	-0.88
Trigger 3	Adverse drug reaction	2	6	92	-0.03 (-0.06-0.002)	0.29	0.84(0.68-0.93)	-0.04	-0.92
Trigger 4	Unplanned transfer to the ICU	15	18	93	0.75 (0.57-0.92)	0.45	0.86 (0.70-0.94)	-0.03	-0.67
Trigger 5	Unplanned return to the operating room	8	12	96	0.78 (0.59-0.99)	0.13	0.92 (0.78-0.98)	-0.04	-0.80
Trigger 6	Unplanned removal, injury or repair of an organ during surgery	5	6	95	0.52 (0.15-0.89)	1.00	0.90 (0.75-0.97)	-0.01	-0.89
Trigger 7	Healthcare related infection or sepsis	20	22	88	0.64 (0.45-0.83)	0.77	0.76 (0.58-0.88)	-0.02	-0.58
Trigger 8	Other complications such as CVA/lung embolism/acute myocardial infarction/TIA	17	16	89	0.60 (0.38-0.81)	1.00	0.78 (0.60-0.89)	0.01	-0.67
Trigger 9	Development of neurological deficit which was not present on admission	5	7	92	0.29 (-0.06-0.65)	0.73	0.84 (0.68-0.93)	-0.02	-0.88

Trigger 10	(Initial) unexpected and/or sudden death, (no palliative care)	12	14	84	0.29 (0.03-0.55)	0.80	0.68 (0.49-0.81)	-0.02	-0.74
Trigger 11	Cardiac or respiratory arrest	12	11	93	0.66 (0.42-0.89)	1.00	0.86 (0.70-0.94)	0.01	-0.77
Trigger 12[4]	Injury related to abortion or delivery	--	--	--	--	--	--	--	--
Trigger 13	Dissatisfaction with care documented in the record	2	5	95	0.27 (-0.17-0.71)	0.38	0.90 (0.75-0.97)	-0.03	-0.93
Trigger 14	Documentation indicating a legal claim or complaint procedure[5]	0	2	--	--	--	--	--	--
Trigger 15	Other patient complication	5	27	72	0.04 (-0.10-0.189)	0.00	0.44 (0.23-0.62)	-0.22	-0.68

[1] *Negative is 100-N*

[2] *PABAK: prevalence and bias adjusted kappa*

[3] *A readmission was considered as unplanned if admission was through the emergency department*

[4] *This trigger was not found in both rounds of the medical record analysis*

[5] *No statistics could be computed because trigger 14 is not present in the first round*

Discussion

In this study, we have shown that the reproducibility (Kappa) of the presence of any trigger present in deceased patients in the hospital was 0.5 (95%CI 0.34-0.66). The average Kappa of individual triggers was 0.42 (range 0-0.78). Our average Kappa of 0.5 (moderate agreement according to Landis et al), appears to be slightly lower than results found in other studies, where a range between 0.49-0.76 was reported.[3,4,21,25-28] However, compared to three Dutch reports which included results of screening for triggers in a sample of cases in 21 hospitals, our Kappa was within the same range.[3,4,21] Naessens (2010) and Ock (2015) evaluated the inter-rater reliability for individual triggers selected either from the HMPS study or the IHI trigger system or both. Four of the triggers investigated by Naessens et al, were comparable to our triggers. Half of these had a higher Kappa agreement in our study compared to Naessens et al. Two out of the three comparable triggers in the study of Ock et al (2015) performed better in their study compared to ours.[29,30] However, again, the population here was sampled from living non-pediatric inpatients.

Concerning the average observed agreement for individual triggers, Unbeck et al (2014) reported that the reproducibility of the individual triggers was on average 46%, in comparison with 67% reproducibility in our study. The total agreement for any trigger present was 65.0% compared to 90% in our study.[31] Regrettably, Unbeck et al didn't report the performance of the triggers on an individual level and studied only living pediatric inpatients. Therefore, this is the first study investigating the performance of the individual triggers of the HMPS trigger system solely in deceased patients. Not surprisingly, objective triggers were more reproducible than subjective triggers.

The Kappa coefficient is influenced by the prevalence of the condition and by bias. Therefore, we also calculated the PABAK. This improved the reliability score, resulting in moderate substantial to almost perfect inter-tester reliability for the individual triggers. An exception was the trigger concerning other patient complications which showed almost no improvement with an end result still well below moderate reliability. The outcomes of these calculations suggest that the low value of Kappa was influenced mainly by the low prevalence of triggers.

Obviously, the performance of the trigger system is important. It should not miss records with serious and potentially preventable AEs and preferably not select any records without AEs. Because trigger systems are used as an aid to reduce the burden of scrutinizing all records, it implies that important AEs could be missed. The fact that new cases with triggers were found in the second round supports this idea.

Due to our random sample of records, we believe that the calculation of the estimated sensitivity and specificity approaches reality. Therefore, when we apply these values found in this study to the entire population of deceased patients in our hospital the false negative rate would be 8%.

The high sensitivity of the system to find cases with an AE was rather comforting. In contrast to a high sensitivity, the specificity of this trigger system was rather low (58%) compared to most of the other studies.[22,32-35] This results in a substantial number of cases

that have to be scrutinized without finding an AE. However, equal results were presented by Howard et al (2017) and our results were slightly better in comparison to Neubert et al (2006), Eggleton et al (2014) and Matlow et al (2011).[36-39]

The variability in triggered cases with a low Kappa suggests unfavourable characteristics of this system. This possibly results in considerable useless time-consuming scrutinizing of records by expensive specialists. Solutions to improve efficacy could be the use of more reproducible triggers (such as the objective ones), combining triggers with patient characteristics, or fully computerized trigger detection by "data mining" software.[23] Before implementing such adaptations, we suggest thorough research concerning the exact performance and costs for finding preventable AEs. However, at the moment there are no better systems available for case selection.

Among the strengths of our study is the fact that the nurses were blinded to the results of the first trigger round. Furthermore, in our system, there were no time limitations while searching for triggers. We, therefore, assume that cases were investigated thoroughly and complete which makes the possibility of a missed trigger as low as possible.

A disadvantage of our study is the small randomly selected sample of all cases that were screened previously. However, this sample was strong enough to detect the real proportion of triggers. Another issue could be the selection of deceased patients. Some studies report that a focus on deaths may not be the most efficient approach or an unsuitable indicator to compare the quality of hospitals.[11,40] Yet, mortality is the event caretakers and patients want to prevent. Of course, departments with low mortality or those who treat non-life-threatening diseases, such as ENT and dermatology, will rarely hear about their AEs from this type of medical record review. As several studies show, AEs don't have to result in death.[2,26,41] They can also cause temporary or permanent injury. Triggers are indications for all AEs, not necessarily for those who cause death. Hence, another chart review system could be more applicable to those departments. Finally, trigger 1 was changed during the time course in which we selected cases for this study. This could have potentially influenced the results. However, trigger 1 was found more often in the second screening round where we expected less often because we shortened the time period making it positive. Therefore, we do not think this influenced our results materially.

Interestingly, more triggers were found during the second round, especially trigger 15 was significantly more present being responsible for most of the difference. In our opinion, supported by a p-value <0.0001 resulting from the McNemar's chi-square test this cannot be attributed to chance alone. We suspect that extra attention among the nurses due to the fact that the second round of review was part of a study might have contributed to this. Furthermore, one could suspect an increase in experience although our team of nurses was deployed for many years in a stable team and cases were selected from a recent period. However, some of the cases that were triggered the first time were not found in the second round. Although memorizing the results in a specific case could have given rise to bias, we found only 12 cases that were checked by the same nurse. Excluding these 12 cases did not influence the results significantly.

We realise that we only analysed a small part of the complete process of looking back at

our proceedings, determine essential parts, develop new solutions and applying them in future care. Moreover, there is no information about the performance of this trigger system in improving health care. However, we think it is important to increase knowledge about these components to optimise care in the end.

Conclusion

In conclusion, applying the adjusted HMPS trigger system as an aid to select records of deceased patients with possible AEs has, in our opinion, a suboptimal Kappa possibly influenced by the low prevalence of individual triggers. Awaiting better selection systems this is, however, the best way to avoid doing a time consuming and costly analysis of all cases. Moreover, it can identify possible threats to quality and safety which can then be further investigated by other methods. However, we have to realise that selection of cases for a more thorough investigation by these common trigger systems, with many subjective triggers, is only moderately reproducible with a low specificity for AEs. Therefore, studies to evaluate possible improvements of these systems or even other systems are important to increase the expediency of these costly tools.

References

1. Kohn L, T., Corrigan J.M., Donaldson, M. To Err is human: building a safer health system. Washington, DC: 1999.
2. Zegers M, de Bruijne MC, Wagner C, Hoonhout LH, Waaijman R, Smits M, et al. Adverse events and potentially preventable deaths in Dutch hospitals: results of a retrospective patient record review study. Qual Saf Health Care. 2009;18(4):297-302.
3. Langelaan M. Monitor zorggerelateerde schade 2011/2012 - Dossieronderzoek in Nederlandse ziekenhuizen. EMGO+ Instituut/VUmc, NIVEL, Nederlands Instituut voor onderzoek van de gezondheidszorg, 2013.
4. Langelaan M, Baines, R.J., Broekens, M.A. . Monitor Zorggerelateerde Schade 2008 - Dossieronderzoek in Nederlandse ziekenhuizen. EMGO+ Instituut/VUmc, NIVEL, Nederlands Instituut voor onderzoek van de gezondheidszorg, 2008.
5. Landrigan CP, Parry GJ, Bones CB, Hackbarth AD, Goldmann DA, Sharek PJ. Temporal trends in rates of patient harm resulting from medical care. N Engl J Med. 2010;363(22):2124-34.
6. Woloshynowych M, Neale G, Vincent C. Case record review of adverse events: a new approach. Qual Saf Health Care. 2003;12(6):411-5.
7. Resar RK, Rozich JD, Classen D. Methodology and rationale for the measurement of harm with trigger tools. Qual Saf Health Care. 2003;12 Suppl 2:ii39-45.
8. Brennan TA, Leape LL. Adverse events, negligence in hospitalized patients: results from the Harvard Medical Practice Study. Perspect Healthc Risk Manage. 1991;11(2):2-8.
9. Classen DC LR, Provost L, Griffin FA, Resar R. Development and evaluation of the Institute for Healthcare Improvement Global Trigger Tool. Journal of Patient Safety 2008;4(3):169-77.
10. Unbeck M, Schildmeijer K, Henriksson P, Jurgensen U, Muren O, Nilsson L, et al. Is detection of adverse events affected by record review methodology? an evaluation of the "Harvard Medical Practice Study" method and the "Global Trigger Tool". Patient Saf Surg. 2013;7(1):10.
11. Hogan H, Healey F, Neale G, Thomson R, Vincent C, Black N. Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study. BMJ Qual Saf. 2012;21(9):737-45.
12. Baines RJ, Langelaan M, de Bruijne MC, Asscheman H, Spreeuwenberg P, van de Steeg L, et al. Changes in adverse event rates in hospitals over time: a longitudinal retrospective patient record review study. BMJ Qual Saf. 2013;22(4):290-8.
13. Baines R, Langelaan M, de Bruijne M, Spreeuwenberg P, Wagner C. How effective are patient safety initiatives? A retrospective patient record review study of changes to patient safety over time. BMJ Qual Saf. 2015;24(9):561-71.
14. Baines RJ, Langelaan M, de Bruijne MC, Wagner C. Is researching adverse events in hospital deaths a good way to describe patient safety in hospitals: a retrospective patient record review study. Bmj Open. 2015;5(7).
15. Chen A, Retegan C, Vinluan J, Beiles CB. Potentially preventable deaths in the Victorian Audit of Surgical Mortality. ANZ J Surg. 2017;87(1-2):17-21.
16. Schoeneberg C, Schilling M, Probst T, Lendemans S. Preventable and potentially preventable deaths in severely injured elderly patients: a single-center retrospective data analysis of a German trauma center. World J Surg. 2014;38(12):3125-32.
17. Lau H, Litman KC. Saving lives by studying deaths: using standardized mortality reviews to improve inpatient safety. Jt Comm J Qual Patient Saf. 2011;37(9):400-8.
18. Zimmerman R, Pierson S, McLean R, McAlpine SA, Caron C, Beth Morrismorris B, et al. Aiming for zero preventable deaths: using death review to improve care and reduce harm. Healthc Q. 2010;13 Spec No:81-7.
19. Sharek PJ. The Emergence of the Trigger Tool as the Premier Measurement Strategy for Patient Safety. AHRQ WebM&M. 2012;2012(5).
20. Langelaan M, Bruijne de, M. C., Baines, R.J., . Monitor zorggerelateerde schade 2011/2012 - Dossiersonderzoek in nederlandse ziekenhuizen 2013.
21. de Bruine MC, Zegers, M., Hoonhout, L.H.F., Wagner, C.,. Onbedoelde schade in Nederlandse ziekenhuizen - Dossieronderzoek van ziekenhuisopnames in 2004 EMGO+ Instituut/VUmc, NIVEL, Nederlands Instituut voor onderzoek van de gezondheidszorg. 2004.
22. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. N Engl J Med. 1991;324(6):370-6.
23. Klein DO, Rennenberg R, Koopmans RP, Prins MH. The ability of triggers to retrospectively predict potentially preventable adverse events in a sample of deceased patients. Prev Med Rep. 2017;8:250-5.
24. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1993;46(5):423-9.
25. Sari AB, Sheldon TA, Cracknell A, Turnbull A. Sensitivity of routine system for reporting patient safety incidents in an NHS hospital: retrospective patient case note review. BMJ. 2007;334(7584):79.

26. Soop M, Fryksmark U, Koster M, Haglund B. The incidence of adverse events in Swedish hospitals: a retrospective medical record review study. Int J Qual Health Care. 2009;21(4):285-91.
27. Wilson RM, Michel P, Olsen S, Gibberd RW, Vincent C, El-Assady R, et al. Patient safety in developing countries: retrospective estimation of scale and nature of harm to patients in hospital. BMJ. 2012;344:e832.
28. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.
29. Naessens JM, O'Byrne TJ, Johnson MG, Vansuch MB, McGlone CM, Huddleston JM. Measuring hospital adverse events: assessing inter-rater reliability and trigger performance of the Global Trigger Tool. Int J Qual Health Care. 2010;22(4):266-74.
30. Ock M, Lee SI, Jo MW, Lee JY, Kim SH. Assessing Reliability of Medical Record Reviews for the Detection of Hospital Adverse Events. J Prev Med Public Health. 2015;48(5):239-48.
31. Unbeck M, Lindemalm S, Nydert P, Ygge BM, Nylen U, Berglund C, et al. Validation of triggers and development of a pediatric trigger tool to identify adverse events. BMC Health Serv Res. 2014;14:655.
32. Classen DC, Resar R, Griffin F, Federico F, Frankel T, Kimmel N, et al. 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. Health Aff (Millwood). 2011;30(4):581-9.
33. Sharek PJ, Parry G, Goldmann D, Bones K, Hackbarth A, Resar R, et al. Performance characteristics of a methodology to quantify adverse events over time in hospitalized patients. Health Serv Res. 2011;46(2):654-78.
34. Landrigan CP, Stockwell D, Toomey SL, Loren S, Tracy M, Jang J, et al. Performance of the Global Assessment of Pediatric Patient Safety (GAPPS) Tool. Pediatrics. 2016;137(6).
35. Lander L, Roberson DW, Plummer KM, Forbes PW, Healy GB, Shah RK. A trigger tool fails to identify serious errors and adverse events in pediatric otolaryngology. Otolaryngol Head Neck Surg. 2010;143(4):480-6.
36. Matlow AG, Cronin CM, Flintoft V, Nijssen-Jordan C, Fleming M, Brady-Fryer B, et al. Description of the development and validation of the Canadian Paediatric Trigger Tool. BMJ Qual Saf. 2011;20(5):416-23.
37. Howard IL, Bowen JM, Al Shaikh LAH, Mate KS, Owen RC, Williams DM. Development of a trigger tool to identify adverse events and harm in Emergency Medical Services. Emerg Med J. 2017;34(6):391-7.
38. Neubert A, Dormann H, Weiss J, Criegee-Rieck M, Ackermann A, Levy M, et al. Are computerised monitoring systems of value to improve pharmacovigilance in paediatric patients? Eur J Clin Pharmacol. 2006;62(11):959-65.
39. Eggleton KS, Dovey SM. Using triggers in primary care patient records to flag increased adverse event risk and measure patient safety at clinic level. N Z Med J. 2014;127(1390):45-52.
40. Hogan H, Zipfel R, Neuburger J, Hutchings A, Darzi A, Black N. Avoidability of hospital deaths and association with hospital-wide mortality ratios: retrospective case record review and regression analysis. BMJ. 2015;351:h3239.
41. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, et al. Incidence of adverse events and negligence in hospitalized patients: results of the Harvard Medical Practice Study I. 1991. Qual Saf Health Care. 2004;13(2):145-51; discussion 51-2.

Chapter 5

Adverse event detection by medical record review is reproducible, but the assessment of their preventability is not

Klein DO, Rennenberg RJMW, Koopmans RP, Prins MH

PLoS ONE 2018; 13(11): e0208087

Abstract

Introduction: The main question in this study was: ''What is the reproducibility of the judgment of medical records by an internal committee reviewing medical records of deceased patients regarding AE presence. Moreover, we also evaluate the root cause of AEs and their reproducibility of these AEs concerning their preventability and their contribution to death?''

Methods: Reviewers re-examined fifty medical records of deceased patients regarding the presence of AE, their potential preventability and their possible contribution to death. Also we investigated the root causes of the preventable AEs.

Results: The Kappa on the presence of an AE was 0.64 and 0.32 for the potential preventability. The intrarater agreement showed a Kappa of 0.61 on the AE presence and 0.64 for the potential preventability. Interrater agreement showed a Kappa of 0.66 for the AE presence and 0.03 for the potential preventability.

Conclusion: We found a fair reproducibility for the detection of AEs, but a poor reproducibility for the potential preventability. Possibly this was caused by lack of a definition for the preventability of AEs. To our opinion an international consensus on what exactly constitutes preventability of AEs and agreement on a definition is necessary.

Introduction

The quality and safety of patient care have gained attention in the past decades. Many methods are used with the aim to improve the quality of care. One of these methods is medical record review.

In several university hospitals in the Netherlands the focus of medical record review is on the detection of preventable adverse events (AEs) in patients who have died during their stay. The records of this group of patients are assumed to contain more (preventable) AEs in comparison with discharged (alive) patients.[1]

Often, trigger systems are used for medical record review because these lower the burden of review, by selecting records. Triggers are clues to alert the screeners for potential AEs so the medical record can be reviewed to determine if an actual AE has occurred.[2-5] Nonetheless, the conclusion that a potentially preventable AE has occurred is stressful for the involved professionals. A common reaction is therefore to dispute the judgment of the committee instead of evaluating the case itself and subsequently improving care.[6,7] To prevent discussions regarding the accuracy of the judgment post hoc, this judgment should be reliable and reproducible.[8-13]

It is likely that if the committee judgment would prove to be reproducible, professionals are more inclined to accept the outcome. There are several studies which investigated the inter-rater reliability using the Harvard medical practice study (HMPS) trigger tool regarding the presence and the preventability of an AE in a mixed (alive and deceased) population. However, these studies showed a moderate reliability for the presence and preventability of an AE.[14-29] Moreover, improving quality and safety of care in an already reasonable safe environment leads to an exponential increase in costs. Therefore, we think health care providers should select the most optimal tests to measure and improve their performance. In order to determine the best methods, qualifying them, making them comparable, and optimising their use, further research is necessary. In our search for better, more reliable methods we hypothesized that our method would show better outcomes compared to previous studies because in our centre the entire review committee discusses the results of the medical record evaluation instead of in pairs or by oneself as is usually reported.[4,30-32]

Therefore, our main question in this study was: ''What is the reproducibility of the judgment of medical records by an internal committee reviewing medical records of deceased patients regarding AE presence. Moreover, we also evaluate the root cause of AEs and their reproducibility of these AEs concerning their preventability and their contribution to death?''

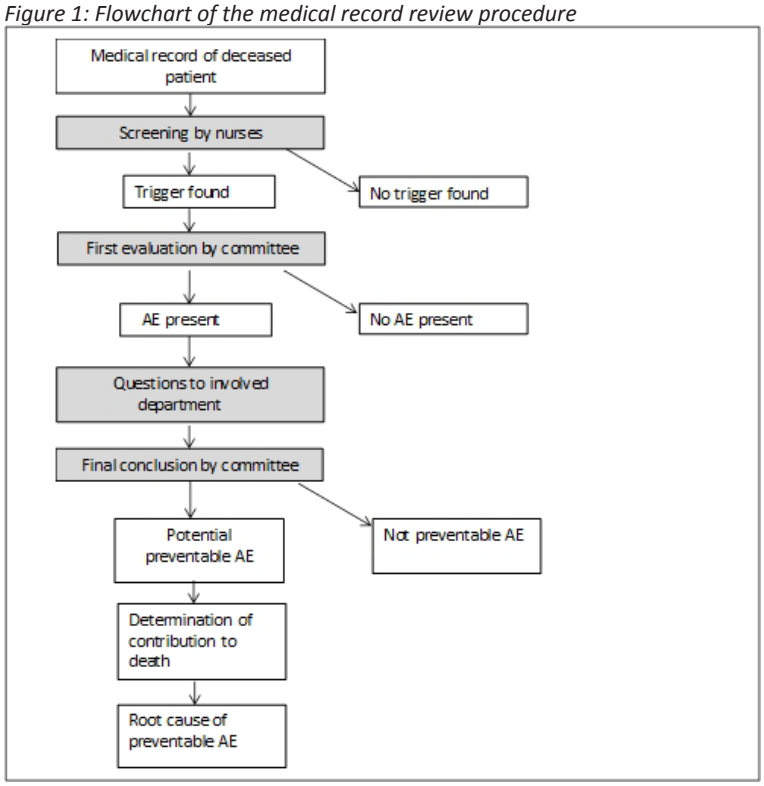
Methods

Medical record review method

Since 2008, a stable team of experienced nurses (each with more than 10 years of clinical experience) with an affinity for record review and patient safety, screen the medical records of all deceased patients for the presence of triggers. Medical record review was introduced in our hospital in 2008 as a project called ''preventing medical injury.'' The aim of this project was to improve the quality and safety of patient care by reducing and prevention of unintended medical injury happening to patients in our hospital.

We use a slightly adapted version of the HMPS trigger system16 to make it suitable for the screening of deceased patients. Therefore, triggers regarding transfer to another acute care hospital and unplanned inappropriate discharge to home were omitted, as they have no relevance in deceased patients. In case a medical record has at least one trigger, it is forwarded to the review committee. Further explanation on the trigger system is published in larger detail elsewhere.[33] The procedure of the medical record review is also shown in figure 1. The study protocol was approved by the medical ethics committee of our hospital. The committee explicitly gave their oral consent to participate in this study and publication of the results.

Figure 1: Flowchart of the medical record review procedure



Definitions

The following definitions were applied in the medical record review:
An AE was defined as an unintended outcome caused by the (non-)action of a caregiver and/ or the healthcare system resulting in temporary or permanent disability or death of the patient. An AE was considered as preventable, if in retrospect after a systematic analysis of the events, it seems that certain measures could have resulted in the prevention of the AE.[34]

Committee

The medical records with a trigger which are forwarded to the review committee are redirected to the member with experience and expertise concerning the case. For example, if a patient died during a surgical procedure, a surgeon analyses the medical record for AEs. We chose this approach because we believe that the considerations on the presence of an AE are best made by a specialist experienced with the usual procedures, protocols and the possible treatments for a certain disease. The committee consists of 9 specialists with different specialties (internist, neurologist, pulmonologist, pediatrician/neonatologist, internist/oncologist, anaesthetist, surgeon, cardiologist, cardiothoracic surgeon and internist/geriatrician) representing the departments with most of the in-hospital deaths. The members are both active (working in our hospital) and recently retired (who used to work in our hospital) experienced specialists.

After a thorough investigation by this committee member, the results are presented to the rest of the committee. A first conclusion on the potential presence of an AE is established in a weekly meeting with at least 4 committee members present. Subsequently, after consulting the involved physicians, the committee finally decides on the presence of an AE, the potential preventability, and the possible contribution of this AE to the death of the patient. If an AE was considered potentially preventable the committee identified the preventable cause according to a standard list of factors (supplementary data, table 1).

Data

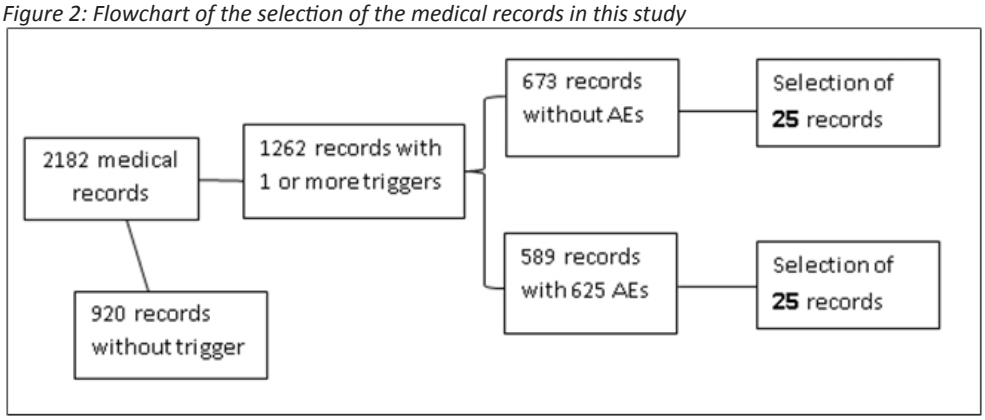
The results of the triggering by the nurses and the review by the specialists are saved using software provided by Medirede®, Clinical File Search version 3 (Mediround BV, 2015). This software was designed to store the data of the medical record review in a clear and easily accessible way.

Selection of records for this study

With regard to our main research question we aimed to get a point estimate with a 95% confidence interval of 6% to each side. Hence, we needed to include 50 cases for the evaluation of the reproducibility of the AE. We randomly selected these medical records out of the database which contained 1262 previously triggered and evaluated medical records of deceased patients between 2013 and 2015. This period was chosen to avoid the presence of recently debated cases in the study sample and therefore minimizing the risk of recognition of cases and remembering former decisions by the committee members. Fig 1 shows the random selection procedure. We used the Excel random generator for the random selection of the study records. Of the 50 randomly selected cases, 25 had an AE after the first evaluation by the committee. We blinded the committee for the previous results of their medical record review. Therefore, the digital results of the first screening

round were inaccessible and thus irretrievable for the reviewers.

Figure 2: Flowchart of the selection of the medical records in this study



Analysis

After this re-evaluation of the medical records by the committee, we compared the outcomes of the two rounds regarding the presence of AEs, the preventability and their contribution to the death of the patient on a team level. The preventability of an AE was scored as being potentially preventable or not preventable. The contribution of an AE to the death of the patient was also scored dichotomous as potential contributing to death or no contribution.

Hence, we calculated the observed overall agreement (in %) to get an impression of the reproducibility of the judgment. By using cross tabulation, we calculated the observed overall agreement (accuracy) within the three groups (presence, preventability, and contribution) with the corresponding 95% confidence interval. Therefore, the number of both rounds negative (without AE) and both rounds positive (with AE) was summed and divided by the total number of medical records. Furthermore, we calculated the positive and negative agreement separately. To determine the reliability for AE presence, preventability, and contribution to death, we executed Cohen's Kappa statistics and the McNemar's test. A p-value of <0.05 was considered statistically significant. Because the judgment of the committee might be influenced by the presenter of a case, we also evaluated the outcomes of the medical record review after the review and presentation to the committee by different specialists (inter-rater reliability) and by the same specialist (intra-rater reliability). Thereafter, we checked if the AEs found in the second evaluation were the same as the original ones found. Finally, we compared the preventable causes to determine agreement between the first and second evaluation.

Results

Analyses comparing the outcomes of the first and the second round on a team level

Presence of an AE (Table 1a)

During the first screening of the medical records (previous to the study), the committee found 25 AEs, the second time (during the study) 28 AEs were found. There was an overlap of 22 AEs which were found both times. The observed agreement for the presence or absence of an AE was 82.0%. The corresponding Cohen's Kappa was 0.64 (95%CI 0.48-0.80). McNemar's test showed a value of 0.51, this means there was no significant difference in the proportion of AEs found in both rounds.

Table 1a: AEs found in first and second round

		Second round		
		Absent	Present	Total
First round	Absent	19	6	25
	Present	3	22	25
Total		22	28	50

Presence of potential preventability

In the first round, 17 of the 25 AEs had the indication to be potentially preventable, in the second round 11 of the 28 AEs were found to be potentially preventable. The corresponding observed overall agreement of the preventable AES was 65% (Table 1b), and the Cohen's Kappa was 0.32 (95%CI 0-0.65). McNemar's test showed a value of 0.070, which indicates there was no significant difference in the proportion of preventability found in both rounds.

Table 1b: Potential preventability of the AEs found

		Second round		
		Not preventable	Potentially preventable	Total
First round	Not preventable	5	1	6
	Potentially preventable	7	10	17
Total		12	11	23

Possible contribution to death of the patient

For the calculation of the agreement regarding the possible contribution of the AE to the death of the patient, we compared the cases (22) with the same AEs in both rounds. 21 of these AEs were considered to possibly have contributed to the death of the patient, in the first round. (McNemar=0.125) During the second screening, the committee concluded that 22 of the 22 AEs possibly contributed to the death of the patient. The accuracy was, therefore, 95%.

Table 1c: Potential contribution to death in both rounds

		Second round		
		No contribution	Potential contribution	Total
First round	No contribution	0	1	1
	Potential contribution	0	21	21
Total		0	22	22

Intrarater reliability – repeatability (individual level)

31 of the 50 cases were reviewed by the same committee member during the evaluation in both rounds. The observed overall agreement for the presence of an AE in this subgroup was 81%, and the corresponding Kappa was 0.61 (95%CI:0.33-0.89). The agreement on the presence of an AE was 79% and the agreement on the absence of an AE was 82%. McNemar's test showed a value of 1.000, thus there was no significant difference found in the proportion of AEs in both rounds. For the preventability of the AEs the observed overall agreement was 57% and the corresponding Kappa 0.64 (95%CI:0.18-1). The agreement on the presence of an AE was 100% and the agreement on the absence of an AE was 60%. An exact McNemar's test determined that there was no significant difference in terms of the preventability of the AEs found the first and second time (p=0.500). 11 cases in which a possible contribution to death was concluded in the first screening, the second screening found the same result, this means a 100% agreement.

Interrater reliability – reproducibility (individual level)

19 of the 50 medical records were scrutinized by different doctors during their evaluation in both rounds. The observed overall agreement for the presence of an AE was 84%, and the Cohen's Kappa was 0.66 (95%CI:0.30-1) in this subgroup. The agreement on the presence of an AE was 79% and the agreement on the absence of an AE was 100%. McNemar's test showed a p-value of 0.250, no significant difference in the proportion of AEs found in both rounds. For the preventability, the observed overall agreement was 45%, and the Cohen's Kappa was 0.03 (95%CI:0-0.55). The agreement on the presence of an AE was 75% and the agreement on the absence of an AE was 29%. An exact McNemar's test determined that there was no significant difference in the proportion of preventability of the AEs found the first and second time (p=0.219). In 10 out of the 11 cases in which a possible contribution to the death of the patient was determined in the first screening, the second screening concluded the same, resulting in 91% agreement.

Root cause analysis

The total number of cases with a potentially preventable adverse event in both sessions, hence labeled with a suspected cause, was nine. The overall agreement on this cause was 78%, with a Kappa of 0.5 (95%CI:0-1). McNemar's test was 1.00 indicating no significant difference.

Discussion

This is the first study assessing the inter- and intrarater reliability for the presence of AEs, their preventability and potential contribution to death among deceased hospital patients. We showed that independent of the committee member involved in the second evaluation, there is a good reliability for the presence of an AE, but only a fair reliability regarding the preventability of an AE. Potential contribution to death was highly reproducible. However, it should be realised that this finding could be distorted by the small number of cases (11) with the same AE in both rounds which were considered to contribute to the death of the patient. Moreover, there was almost no AE that did not contribute to the death of a patient. At the same time, it was difficult to exclude that an AE did not contribute to the death of the patient. It seems, therefore, that the trigger system especially selects cases with AEs that contribute to or cause death. In addition, we couldn't calculate the Kappa for this section due to the low numbers in the cells of the two by two tables.
Our committee found more AEs when scrutinizing the same records for the second time. Therefore, the question remains if the number of AEs found eventually is the true number of AEs. Even if the inter-rater reliability would be acceptable, there is no evidence that this kind of record review really detects all AEs.[14]

A strong point of our study is, in contrast to most other studies, the fact that reviewers discussed the outcomes of the medical record analyses in a group session. In these other studies, usually pairs of reviewers perform the medical record review. A study by Hofer et al (2000) found that discussion between reviewers didn't improve the overall reliability, with a Kappa of 0.36 before and a Kappa of 0.40 after discussion.[35] Interestingly, in our study, in which we discussed cases twice (during the first and second round), we found a substantially higher level of agreement with a Kappa of 0.64. However, the discussion in our study was between the same committee members both times, whereas in the study of Hofer et al, it was between different pairs of reviewers.

The results from our subgroup analyses suggest a rather low inter-rater reproducibility concerning the preventability of the AEs, in contrast to a higher intra-rater reproducibility. The same phenomenon is also seen in other studies in which both hospital survivors and deceased patients were included.[23,25,26,28-30,36,37] Although our system is different in some crucial parts, the judgement concerning the preventability of AEs is comparable with the results in these studies.

Several studies use different classifications for preventability. 3-level classification systems and also a 5- and 6- level classification systems are used in studies.[23,28,29,36] This makes comparisons between studies difficult. There is no gold standard concerning potential preventability which leaves us with the consensus of the medical record research committee as second best. Differences in opinion will therefore always exist, this might give rise to an ongoing discussion. Nevertheless, the effort to detect AEs and their preventability is in our opinion useful because any preventable cause that can be abolished is important, therefore adequate feedback, resulting in consecutive change of procedures if necessary, is needed.

Our study had some limitations. A few methodological aspects of our study require attention. We blinded our reviewers and used records from several years ago, which we consider as a strong point, although we cannot exclude that they may have been able to recognize cases from their memory. Furthermore, there is the lack of a clear definition (which is also not available in international literature) to determine preventability. We think this is the reason why the reliability of the judgment on preventability is disappointing. Although the committee judgment on preventability became based on more experience over the past years, there are no strict guidelines or rules on how to judge. Previous studies sometimes use Likert scales or percentages which in our opinion creates a false sense of precision, with inconsistent reproducibility.[13,20,38-42] Secondly, the sample size of the re-examined cases was rather small, especially in the subgroup analysis. This was reflected in the root cause analysis in which only nine cases remained to be analysed. Therefore, we were unable to answer all of our sub questions, especially the one focusing on the reproducibility of the contribution of the AE to the death of the patient. The small number of cases was the result of powering the study to our main question. Moreover, increasing the number of cases to answer our sub-questions with sufficient power would have been very time and hence cost consuming.

In conclusion, we showed that the judgment on the presence of an AE by a committee investigating medical records of deceased patients was reproducible. However, their judgment on preventability is less so, possibly because there is a lack of clear definitions on this subject. It is difficult to comment on the reliability of contribution to death because all AEs seem somehow to contribute to the death of the patient in both rounds. Despite this, we think giving feedback to professionals based on the review of preventable AEs is mandatory. Yet, we should focus on finding more reliable methods to identify preventable AEs. The next step should be a clear international accepted definition on preventability which could aid to increase the reproducibility of these methods and thus make them more valuable. Thereafter, respectful communication with the involved professionals in order to improve the quality and safety of care is of utmost importance to increase the chance that care will really become better.

References

1. Baines RJ, Langelaan M, de Bruijne MC, Wagner C. Is researching adverse events in hospital deaths a good way to describe patient safety in hospitals: a retrospective patient record review study. Bmj Open. 2015;5(7).
2. Brennan TA, Leape LL. Adverse events, negligence in hospitalized patients: results from the Harvard Medical Practice Study. Perspect Healthc Risk Manage. 1991;11(2):2-8.
3. Classen DC LR, Provost L, Griffin FA, Resar R. Development and evaluation of the Institute for Healthcare Improvement Global Trigger Tool. Journal of Patient Safety 2008;4(3):169-77.
4. de Wet C, Bowie P. The preliminary development and testing of a global trigger tool to detect error and patient harm in primary-care records. Postgrad Med J. 2009;85(1002):176-80.
5. Resar RK, Rozich JD, Classen D. Methodology and rationale for the measurement of harm with trigger tools. Qual Saf Health Care. 2003;12 Suppl 2:ii39-45.
6. Millwood S. Developing a Platform for Learning from Mistakes: changing the culture of patient safety amongst junior doctors. BMJ Qual Improv Rep. 2014;3(1).
7. Wolf ZR, Hughes RG. Error Reporting and Disclosure. In: Hughes RG, editor. Patient Safety and Quality: An Evidence-Based Handbook for Nurses. Advances in Patient Safety. Rockville (MD)2008.
8. Contreary K, Collins A, Rich EC. Barriers to evidence-based physician decision-making at the point of care: a narrative literature review. J Comp Eff Res. 2016.
9. He L, Gannon S, Shannon CN, Rocque BG, Riva-Cambrin J, Naftel RP. Surgeon interrater reliability in the endoscopic assessment of cistern scarring and aqueduct patency. J Neurosurg Pediatr. 2016;18(3):320-4.
10. Mooney MA, Hardesty DA, Sheehy JP, Bird R, Chapple K, White WL, et al. Interrater and intrarater reliability of the Knosp scale for pituitary adenoma grading. J Neurosurg. 2016:1-6.
11. Ramnarayan P, Kapoor RR, Coren M, Nanduri V, Tomlinson AL, Taylor PM, et al. Measuring the impact of diagnostic decision support on the quality of clinical decision making: development of a reliable and valid composite score. J Am Med Inform Assoc. 2003;10(6):563-72.
12. Coutts SB, Simon JE, Tomanek AI, Barber PA, Chan J, Hudon ME, et al. Reliability of assessing percentage of diffusion-perfusion mismatch. Stroke. 2003;34(7):1681-3.
13. Sharek PJ, Parry G, Goldmann D, Bones K, Hackbarth A, Resar R, et al. Performance characteristics of a methodology to quantify adverse events over time in hospitalized patients. Health Serv Res. 2011;46(2):654-78.
14. Hanskamp-Sebregts M, Zegers M, Vincent C, van Gurp PJ, de Vet HC, Wollersheim H. Measurement of patient safety: a systematic review of the reliability and validity of adverse event detection with record review. BMJ Open. 2016;6(8):e011078.
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.
16. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. N Engl J Med. 1991;324(6):370-6.
17. Sari AB, Sheldon TA, Cracknell A, Turnbull A. Sensitivity of routine system for reporting patient safety incidents in an NHS hospital: retrospective patient case note review. BMJ. 2007;334(7584):79.
18. Brennan TA, Localio RJ, Laird NL. Reliability and validity of judgments concerning adverse events suffered by hospitalized patients. Med Care. 1989;27(12):1148-58.
19. Baines RJ, Langelaan M, de Bruijne MC, Asscheman H, Spreeuwenberg P, van de Steeg L, et al. Changes in adverse event rates in hospitals over time: a longitudinal retrospective patient record review study. BMJ Qual Saf. 2013;22(4):290-8.
20. Zegers M, de Bruijne MC, Wagner C, Groenewegen PP, van der Wal G, de Vet HC. The inter-rater agreement of retrospective assessments of adverse events does not improve with two reviewers per patient record. J Clin Epidemiol. 2010;63(1):94-102.
21. Thomas EJ, Studdert DM, Burstin HR, Orav EJ, Zeena T, Williams EJ, et al. Incidence and types of adverse events and negligent care in Utah and Colorado. Medical care. 2000;38(3):261-71.
22. Localio AR, Weaver SL, Landis JR, Lawthers AG, Brenhan TA, Hebert L, et al. Identifying adverse events caused by medical care: degree of physician agreement in a retrospective chart review. Annals of internal medicine. 1996;125(6):457-64.
23. Hogan H, Healey F, Neale G, Thomson R, Vincent C, Black N. Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study. BMJ Qual Saf. 2012;21(9):737-45.
24. Soop M, Fryksmark U, Koster M, Haglund B. The incidence of adverse events in Swedish hospitals: A retrospective medical record review study. International Journal for Quality in Health Care. 2009;21(4):285-91.
25. Forster AJ, Asmis TR, Clark HD, Al Saied G, Code CC, Caughey SC, et al. Ottawa Hospital Patient Safety Study: incidence and timing of adverse events in patients admitted to a Canadian teaching hospital. CMAJ. 2004;170(8):1235-40.

26. Baker GR, Norton PG, Flintoft V, Blais R, Brown A, Cox J, et al. The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. CMAJ. 2004;170(11):1678-86.
27. Davis P. LYR, Briant, R. Adverse events in New Zealand public hospitals: principal findings from a national survey. New Zealand: Ministry of Health 2001.
28. Thomas EJ, Lipsitz SR, Studdert DM, Brennan TA. The reliability of medical record review for estimating adverse event rates. Ann Intern Med. 2002;136(11):812-6.
29. Wilson RM, Runciman WB, Gibberd RW, Harrison BT, Newby L, Hamilton JD. The Quality in Australian Health Care Study. Med J Aust. 1995;163(9):458-71.
30. Mortaro A, Moretti F, Pascu D, Tessari L, Tardivo S, Pancheri S, et al. Adverse Events Detection Through Global Trigger Tool Methodology: Results From a 5-Year Study in an Italian Hospital and Opportunities to Improve Inter-rater Reliability. J Patient Saf. 2017.
31. Rutberg H, Borgstedt Risberg M, Sjodahl R, Nordqvist P, Valter L, Nilsson L. Characterisations of adverse events detected in a university hospital: a 4-year study using the Global Trigger Tool method. Bmj Open. 2014;4(5):e004879.
32. Najjar S, Hamdan M, Euwema MC, Vleugels A, Sermeus W, Massoud R, et al. The Global Trigger Tool shows that one out of seven patients suffers harm in Palestinian hospitals: challenges for launching a strategic safety plan. Int J Qual Health Care. 2013;25(6):640-7.
33. Klein DO, Rennenberg R, Koopmans RP, Prins MH. The ability of triggers to retrospectively predict potentially preventable adverse events in a sample of deceased patients. Prev Med Rep. 2017;8:250-5.
34. Wagner C vdWG. Voor een goed begrip. Bevordering patiëntveiligheid vraagt om heldere definities [For a good understanding. Improving patient safety requires clear definitions] Med contact. 2005;60:1888-91.
35. Hofer TP, Bernstein SJ, DeMonner S, Hayward RA. Discussion between reviewers does not improve reliability of peer review of hospital quality. Med Care. 2000;38(2):152-61.
36. Soop M, Fryksmark U, Koster M, Haglund B. The incidence of adverse events in Swedish hospitals: a retrospective medical record review study. Int J Qual Health Care. 2009;21(4):285-91.
37. Davis P, Lay-Yee R, Briant R, Ali W, Scott A, Schug S. Adverse events in New Zealand public hospitals I: occurrence and impact. N Z Med J. 2002;115(1167):U271.
38. Landrigan CP, Parry GJ, Bones CB, Hackbarth AD, Goldmann DA, Sharek PJ. Temporal trends in rates of patient harm resulting from medical care. N Engl J Med. 2010;363(22):2124-34.
39. O'Leary KJ, Devisetty VK, Patel AR, Malkenson D, Sama P, Thompson WK, et al. Comparison of traditional trigger tool to data warehouse based screening for identifying hospital adverse events. BMJ Qual Saf. 2013;22(2):130-8.
40. Zwaan L, De Bruijne M, Wagner C, Thijs A, Smits M, Van Der Wal G, et al. Patient record review of the incidence, consequences, and causes of diagnostic adverse events. Archives of internal medicine. 2010;170(12):1015-21.
41. Mirza SK, Deyo RA, Heagerty PJ, Turner JA, Lee LA, Goodkin R. Towards standardized measurement of adverse events in spine surgery: conceptual algorithm and pilot evaluation. BMC Musculoskelet Disord. 2006;7:53.
42. Kessomboon P, Panarunothai S, Wongkanaratanakul P. Detecting adverse events in Thai hospitals using medical record reviews: agreement among reviewers. J Med Assoc Thai. 2005;88(10):1412-8.

Supplementary data

Table 1: Most important factors contributing to preventable AEs in MUMC+

Factors
1) Diagnostics and other actions prior to treatment Knowledge, skills, code of conduct Wrong/missed diagnosis, under- or overdiagnosis and evaluation Wrong indication (beneficial effect of the treatment doesn't compensate for the burden of the treatment; no beneficial effect expected; less invasive treatment has an equal effect) Estimation of the patient capacity/treatment burden incorrect Other….
2) Treatment Knowledge, skills, interpersonal skills Surgical technique (surgery, instrumental intervention) Medication technical Medication (wrong dosage, side effect) Nursing Other…
3) Follow-up process Wrong assessment of the disease severity/symptoms, complications (missing gut-feeling resulting in no action or too late action/inadequate diagnostics) Follow-up of the patient (inaccurate) Follow-up of the patient (incompetent) Other…
4) Communication, cooperation and reporting Communication during transfer from one location to another Communication during transfer between physicians/nurses/paramedics Cooperation of healthcare providers Inadequate reporting Other….
5) Organizational/technical defects Protocols/procedures/organization Equipment/materials No medium care/ incorrect patient placement assignments Other…
6) Disease-related factors Compliance by the patient Disease severity of the patient Comorbidity Other..
7) Other

Chapter 6

Limited external reproducibility restricts the use of medical record review for benchmarking

Klein DO, Rennenberg, RJMW, Gans ROB, Enting RH, Koopmans RP, Prins MH

BMJ Open quality 2019 8(2):e000564

Abstract

Background: Medical record review is used to assess the quality and safety in hospitals. It's increasingly used to compare institutions. Therefore, the external reproducibility should be high. In the current study, we evaluated this external reproducibility for the assessment of an adverse event (AE) in a sample of records from two university medical centres in the Netherlands, using the same review method.

Methods: From both hospitals, 40 medical records were randomly chosen from patient files of deceased patients that had been evaluated in the preceding years by the internal review committees. After reviewing by the external committees, we assessed the overall and Kappa agreement by comparing the results of both review rounds (once by the own internal committee and once by the external committee). This was calculated for: the presence of an AE, preventability and contribution to death.

Results: Kappa for the presence of AEs was moderate (k=0.47). For preventability, the agreement was fair (k=0.39) and poor for contribution to death (k=−0.109).

Conclusion: We still believe that medical record review is suitable for the detection of general issues concerning patient safety. However, based on the outcomes of this study, we would advise to be careful when using medical record review for benchmarking.

Introduction

In many countries worldwide, healthcare inspection increasingly demand information on the quality and safety of patient care in hospitals. Several tools have been implemented by hospitals for the monitoring of their patients' safety.[1,2] A widely used tool is systematic medical record review (MRR). In the Netherlands, hospitals are obliged to either arrange an internal MRR system or take part in a national monitoring program of care related harm (performed every 4 years) executed by the Netherlands Institute for health services research (NIVEL).[3-5]

Hospitals using an MRR system frequently evaluate a subset of records (for example every tenth admission) to lower the burden of MRR or select cases most likely to contain adverse events (AEs) (for example only patients who died during hospitalization). An additional method to lower the burden of MRR for physicians is to use a trigger system, which is executed by nurses in a previously defined set of records. When one or more triggers are found, the record is evaluated by a review committee. The results of this medical record review should be reliable and valid, because the outcome could lead to changes in care for future patients. Therefore, ideally, results must be both internally and externally reproducible. Internal reproducibility is necessary to obtain support for proposed improvements within a given institution. External reproducibility is necessary to compare results across institutions (benchmarking). However, well-defined criteria guiding the reviewer in how to fulfil a good MRR have not been specified clearly in international literature or guidelines.[6-11]

In the current study, we focus on the committee judgment and analyse the external reproducibility of the committee judgment on a sample of records from two university hospitals in the Netherlands, using the same review method.
Moreover, we also evaluate the root cause of potentially preventable AEs and their corresponding reproducibility.

Methods

Selection of records

For both hospitals, 40 medical records were extracted from patient files of deceased patients that had been investigated and completed in the preceding 2 years by the internal committees (2014-2016) (hospital 1: 448 out of 717 records, hospital 1: 379 out of 512 records, figure 1A and 1B). The first step of the selection was: records were selected according to the expertise of the committee members that participated (see section study). Then the sample was randomly chosen from these departments. The selection of the records was executed by DK using the Excel® random generator.

Figure 1A: Selection of records in centre 1

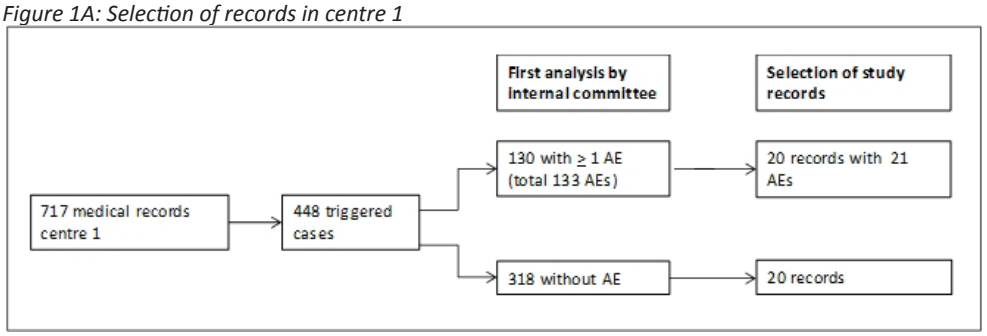
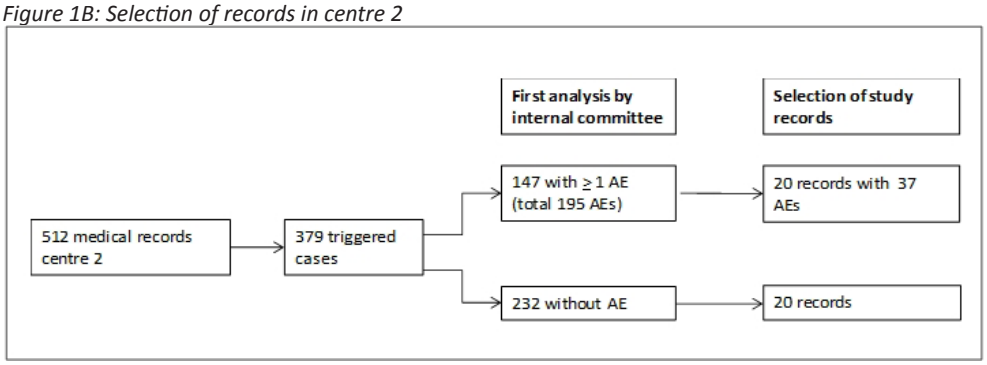


Figure 1B: Selection of records in centre 2



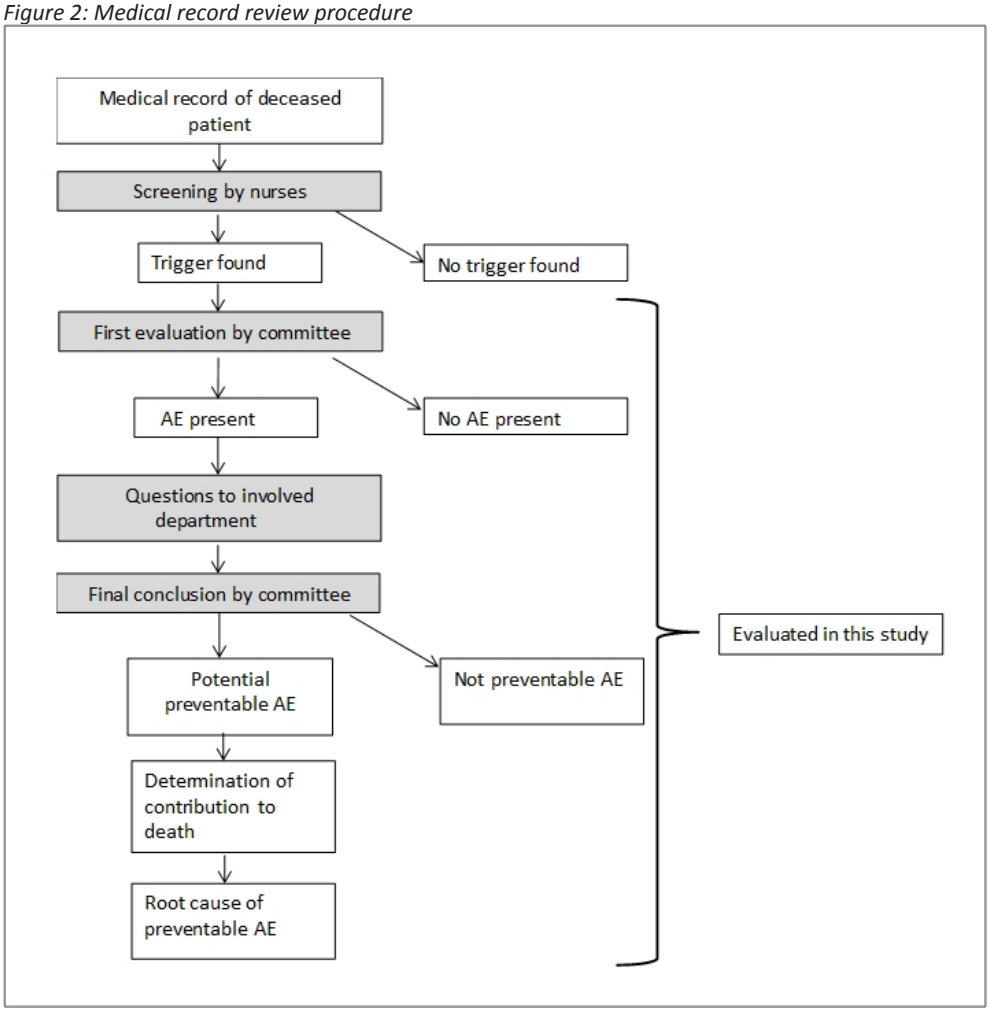
Furthermore, 50% of the records were selected out of the group with suitable records with an AE and the other half out of the group without an AE. Records selected for the committee from hospital one comprise patients originally treated by cardiology, surgery or internal medicine departments. For the committee from hospital two, they were originally treated by internal medicine, surgery, ICU, cardiology or the neurology department. Since we wanted to investigate the external reliability of the review process only, we selected records in which nurses had found triggers when they were evaluated for the first time by the internal committee.

Study

In 2016 we gathered two times for three consecutive days (in 2016) the selected medical records were evaluated on location by the delegates of the external hospital committee. The committee of centre 1 thus evaluated in this study the records of centre 2 and vice versa. The admission department of the patient determined which specialist would (preferably) investigate a specific record. If this would be, for example, surgery, then a surgeon from the committee would evaluate the record. After these three days, the outcome of the evaluation by the delegates was discussed in a consensus meeting in which at least the three delegates were present. This consensus meeting was performed by the committees separately. During this meeting, a conclusion had to be reached on whether an AE had occurred. Furthermore, its potential preventability was assessed and the potential contribution of the AE to the death of the patient was determined. There was no time limitation for the review or the discussion in the committee. Each committee was blinded for the results

of the first evaluation of the records by the other committee. The study process is further clarified in figure 2.[12]

Figure 2: Medical record review procedure



Committees

For hospital 1, the delegates were: an internist, a surgeon, and a cardiologist; there has been an MRR committee in this centre since 2008.

At hospital 2, the delegates were: an internist, cardiothoracic surgeon, and a neurologist; they started with MRR, according to the same format as hospital 1, in 2014. All reviewers took part in the national NIVEL studies and were therefore trained in the same fashion.[5]

During the previous years, both committees used the same review procedure. Previous research showed the results of this internal medical record review to be acceptably reliable.[12]

Training

For the participation in the NIVEL studies, the nurses and physicians followed a one-day training in small groups (max 12 participants) led by one member of the research team and one experienced nurse or physician, respectively. During the training, the study protocol, definitions and review forms were explained and examples of (preventable) adverse events were discussed. The reviewers practiced with cases and they were provided with a review manual. After one month of reviewing, the reviewers had a half-day training session to discuss their problems concerning the review process and definitions and to update the reviewers with the latest insights about the review process. These training sessions were frequently repeated during data collection.

Statistics and analyses

We aimed for a Kappa of 0.6 or more, while we expected a Kappa of 0.75. With a type 1 error of 0.05 and a type 2 error of 0.20, a sample size of 80 cases was found to be sufficient.[13]

To evaluate the output of the external review, we performed the following analyses: overall agreement and corresponding Kappa agreement with a 95% confidence interval. This was executed for the following variables: the presence of an AE, the presence of a potentially preventable AE and the presence of an AE which had contributed to the death of the patient.

By using cross tabulation, we calculated the observed overall agreement (accuracy) within the four groups (presence, preventability, contribution to death and root cause) with the corresponding 95% confidence interval.

Prevalence-adjusted and bias-adjusted kappa (PABAK) calculations were done and reported along with kappa, to show how data would have been with equal distributions of positive and negative test results. Finally, corresponding prevalence and bias indices were calculated.[14]

Furthermore, for every medical record separately, we evaluated the AEs that the committees found. We checked if the same AEs were found as during the first evaluation. If more than one AE was found, we checked if at least the same AE compared to the first evaluation was present. This was also done for preventability of the AEs and the contribution to death of all AEs.

The values of kappa were categorized as follows, the degree of agreement was categorized as poor ($\kappa < 0$), slight (κ=0.00-0.20), fair (κ=0.21-0.40), moderate (κ=0.41-0.60), substantial (κ=0.61-0.80) or almost perfect (κ=0.81-1.00).[15]

Definitions

An AE was defined as an unintended outcome caused by the (non-)action of a caregiver and/ or the healthcare system resulting in temporary or permanent disability or death of the patient.[16]

When an AE had been identified, its potential preventability was assessed (subdivided in the categories not preventable and potentially preventable) and the potential contribution of the AE to the death of the patient was determined (subdivided into: no contribution and potential contribution).

Data storage

All results were saved using software provided by Medirede®, Clinical File Search version 3 (Mediround BV, 2015).

Data safety
The study was approved by the Medical Ethics Committee (of both participating centres). To guarantee patients' privacy, the medical records were only accessible at the centre itself. The selected records were accessible in the digital environment of the hospital. Furthermore, reviewers signed confidentiality contracts.

Results

Eighty records in total were reassessed here we present the results after review by the other committee. Outcomes for all records were available.

Medical records overall agreement

Table 1 shows the evaluation of the cases regarding the presence of an AE. The overall agreement was 74% and the corresponding Kappa 0.48 (95%CI:0.28-0.67). PABAK was 0.48 (95%CI:0.28-0.67).
Table 2 shows of the number of AEs that were found by two teams, the evaluation regarding the potential preventability of this AE. The overall agreement regarding the preventability was therefore 71% and the Kappa agreement 0.39 (95%CI:0.08-0.69). PABAK was 0.41 (95%CI:0.11-0.72)
Table 3 shows the evaluation of both teams regarding the contribution of the AE to death of the patient. The overall agreement regarding this contribution of the AE was 65% and the corresponding Kappa agreement was -0.109 (95%CI:-0.24-0.02). PABAK was 0.29 (95%CI:0-0.61).

Root cause analysis

The total number of cases with a potentially preventable AE according to both committees, hence labeled with a suspected cause, was 4. The overall agreement on this cause was 71%, with a Kappa of 0.481 (95%:CI 0-1).

Table 1: Evaluation of the committees regarding the presence of an AE

	AE present (committee 1)	AE not present (committee 1)	
AE present (committee 2)	34	14	48
AE not present (committee 2)	7	25	32
	41	39	80

Table 2: Evaluation of the committees regarding the potential preventability of the AEs

	AE potentially preventable (committee 1)	AE not preventable (committee 1)	
AE potentially preventable (committee 2)	8	3	11
AE not preventable (committee 2)	7	16	23
	15	19	34

Table 3: Evaluation of the committees regarding the potential contribution of the AEs to death

	AE contributed (committee 1)	AE no contribution (committee 1)	
AE contributed (committee 2)	22	2	24
AE no contribution (committee 2)	10	0	10
	32	2	34

Discussion

This study shows that, although the overall agreement of a judgment seems promising (as shown in table 1) the agreement of the reviewers for the presence of an AE is moderate with a Kappa of 0.47. The agreement for the preventability was fair (K=0.39) and for the contribution of the AEs to death was poor (K=-0.109).17 The calculations of the PABAK show that the prevalence and bias had a negligible effect on the results. Only for the contribution of the AE to the death of the patient, an effect of the prevalence was shown. This indicates that the external reproducibility of medical record review isn't optimal and needs improvement.[18]

The NIVEL studies reported comparable results for the agreement between external reviewers. Their Kappa agreement ranged between 0.24 and 0.47 for the presence of an AE. For preventability of an AE, the kappa was found to be 0.43. The improvement was explained by more intensified training.[3,4]

Sharek et al, and Landrigan et al, also show a moderate agreement for the AE presence and its severity between internal review teams and external review teams. However, the performance of these teams was not evaluated in a second hospital with different cases. This makes a comparison with our study difficult.19,20 Finally, Schildmeijer et al, showed a comparable agreement for the presence of an AE between teams using the GTT method.[21]

Strong points of our study are; the blinding of the two committees for the results of the first review by the other committee. Furthermore, we have chosen two comparable committees from two university hospitals using the same review method, to exclude that the review method itself caused any differences that would be found. Also, this is the first study in which committees of two hospitals review each other's medical records for the evaluation of the external reproducibility. Contrarily to the NIVEL studies, which only compare results of two external committees, we compared the review of an external with an internal committee as is more common in other studies.[19,22] Also, we believe that the reviewers in both committees can be seen as experts, since they evaluate medical records on a regular base (not only for study purposes).23 Furthermore, when we started the study, both teams already performed medical record review for at least three years. The number of records evaluated by these 2 committees per year far exceeded the total number of records in the study by Landrigan et al.[23] This study showed that the agreement improved when the reviewers gained more experience, which we don't think could be the case for our reviewers since they were already experienced at the start of our study. Obviously, there are also some points for improvement.

In our study, we can't exclude differences in the performance of the two committees although both of them apply the same review method. Reasons for this could be as follows. First, the clinical background of the reviewers was slightly different. Second, committee 1 gave their final judgment after consulting other committee members who were not involved in scrutinizing the 40 cases from the external review. Whereas committee 2 recorded the final judgment after reaching consensus in their group of three members. Finally, centre 2 has been active for a shorter period and aims to detect all adverse events, whereas centre 1 with a longer experience focuses on the most severe and preventable AEs. The detection rate of AEs in all records (preventable and not preventable) is therefore much higher in centre 2 than in centre 1 (29% vs. 18%).

Also, the number of records in which the root cause of the AE was noted was too small to draw conclusions on the agreement (this is also reflected by the large confidence interval). Furthermore, committee 1 consisted partly of recently retired specialists while the other committee consisted of solely active physicians. Centre 1 chose to use the expertise of retired specialists since they have more time for the investigation of the records compared to presently active specialists who need to review medical records on top of their usual work. At the same time, in centre 2 the active specialists in the committee get dedicated time for their MRR. Additionally, although the committees were instructed to use their common method for review and final decision we cannot exclude any influence of the fact that the review of the 40 cases in the other hospital was done especially for study purposes. Finally, some of the medical records contained more than 1 AE, which made it easier for the external committee to find at least one of these AE; this could have led to an overestimation of the external reproducibility. Most MRR studies call for more research and exploration of possibilities for improving the inter-rater reliability since there is a need for more good quality studies on this topic.[9,24-28] However, a recent article by Leistikow endorses otherwise. According to this article, the main reason for the disappointing reproducibility of MRR is because it depends on the values and view of the person who is performing the review.[29,30] At the same time, the definitions of an AE and its preventability are changing

over time.[31] Moreover, we should not only apply traditional medical research methods for evaluating patient safety, but also involve behavioural- and social sciences. Organisational behaviour research in healthcare, for example, has highlighted the psychological, social, cultural and economic obstacles to a simple implementation of a solution. These sciences can help in understanding the complexity of patient safety.[32,33] Combining these approaches could provide a better understanding of the complexity of patient safety and help with the design of interventions that are really beneficial for patients.[29]

In conclusion, we think that medical record review is suitable for the detection of general issues in patient safety and also for the discussion of individual cases. However, the suboptimal reproducibility of MRR reduces its potential for benchmarking. Finally, we think at least a better definition of preventability and also of contribution to death is needed if we want to compare the outcomes between hospitals.

References

1. Griffin FA. IHI Global Trigger Tool for Measuring Adverse Events (Second Edition) IHI Innovation Series white paper. Cambridge, Massachusetts: Institute for Healthcare Improvement. 2009.
2. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. N Engl J Med. 1991;324(6):370-6.
3. Zegers M, de Bruijne MC, Wagner C, Hoonhout LH, Waaijman R, Smits M, et al. Adverse events and potentially preventable deaths in Dutch hospitals: results of a retrospective patient record review study. Qual Saf Health Care. 2009;18(4):297-302.
4. Baines RJ, Langelaan M, de Bruijne MC, Asscheman H, Spreeuwenberg P, van de Steeg L, et al. Changes in adverse event rates in hospitals over time: a longitudinal retrospective patient record review study. BMJ Qual Saf. 2013;22(4):290-8.
5. Zegers M, de Bruijne MC, Wagner C, Groenewegen PP, Waaijman R, van der Wal G. Design of a retrospective patient record study on the occurrence of adverse events among patients in Dutch hospitals. BMC Health Serv Res. 2007;7:27.
6. Nabhan M, Elraiyah T, Brown DR, Dilling J, LeBlanc A, Montori VM, et al. What is preventable harm in healthcare? A systematic review of definitions. BMC Health Serv Res. 2012;12:128.
7. Weingart SN. Finding common ground in the measurement of adverse events. Int J Qual Health Care. 2000;12(5):363-5.
8. Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. J Biomed Inform. 2003;36(1-2):131-43.
9. Unbeck M, Schildmeijer K, Henriksson P, Jurgensen U, Muren O, Nilsson L, et al. Is detection of adverse events affected by record review methodology? an evaluation of the "Harvard Medical Practice Study" method and the "Global Trigger Tool". Patient Saf Surg. 2013;7(1):10.
10. Walshe K. Adverse events in health care: issues in measurement. Qual Health Care. 2000;9(1):47-52.
11. Jha AK, Classen DC. Getting moving on patient safety--harnessing electronic data for safer care. N Engl J Med. 2011;365(19):1756-8.
12. Klein DO, Rennenberg R, Koopmans RP, Prins MH. Adverse event detection by medical record review is reproducible, but the assessment of their preventability is not. PLoS One. 2018;13(11):e0208087.
13. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med. 1998;17(1):101-10.
14. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1993;46(5):423-9.
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.
16. Wagner C vdWG. Voor een goed begrip. Bevordering patiëntveiligheid vraagt om heldere definities [For a good understanding. Improving patient safety requires clear definitions] Med contact. 2005;60:1888-91.
17. Monto AS, Dickson CB, Landis JR. Utilization and acceptability of influenza A/New Jersey/76 virus vaccine in Oakland County, Michigan. J Infect Dis. 1977;136 Suppl:S693-8.
18. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-82.
19. Sharek PJ, Parry G, Goldmann D, Bones K, Hackbarth A, Resar R, et al. Performance characteristics of a methodology to quantify adverse events over time in hospitalized patients. Health Serv Res. 2011;46(2):654-78.
20. Landrigan CP, Parry GJ, Bones CB, Hackbarth AD, Goldmann DA, Sharek PJ. Temporal trends in rates of patient harm resulting from medical care. N Engl J Med. 2010;363(22):2124-34.
21. Schildmeijer K, Nilsson L, Arestedt K, Perk J. Assessment of adverse events in medical care: lack of consistency between experienced teams using the global trigger tool. BMJ Qual Saf. 2012;21(4):307-14.
22. Schildmeijer KG, Nilsson L, Arestedt K, Perk J. The assessment of adverse events in medical care; lack of consistency between experienced teams using the Global Trigger Tool'. BMJ Qual Saf. 2013;22(3):271-2.
23. Landrigan CP, Stockwell D, Toomey SL, Loren S, Tracy M, Jang J, et al. Performance of the Global Assessment of Pediatric Patient Safety (GAPPS) Tool. Pediatrics. 2016;137(6).
24. Hanskamp-Sebregts M, Zegers M, Vincent C, van Gurp PJ, de Vet HC, Wollersheim H. Measurement of patient safety: a systematic review of the reliability and validity of adverse event detection with record review. BMJ Open. 2016;6(8):e011078.
25. Zegers M, Hesselink G, Geense W, Vincent C, Wollersheim H. Evidence-based interventions to reduce adverse events in hospitals: a systematic review of systematic reviews. Bmj Open. 2016;6(9):e012555.
26. Hofer TP, Bernstein SJ, DeMonner S, Hayward RA. Discussion between reviewers does not improve reliability of peer review of hospital quality. Med Care. 2000;38(2):152-61.
27. Farup PG. Are measurements of patient safety culture and adverse events valid and reliable? Results from a cross sectional study. BMC health services research. 2015;15:186.

28. Mattsson TO, Knudsen JL, Lauritsen J, Brixen K, Herrstedt J. Assessment of the global trigger tool to measure, monitor and evaluate patient safety in cancer patients: reliability concerns are raised. BMJ quality & safety. 2013;22(7):571-9.
29. Leistikow I. Aantonen patiëntveiligheid vergt acceptatie van breder wetenschapspalet. Ned Tijdschr Geneeskd 2017;. 2017;161(D1121).
30. Leistikow I, Mulder S, Vesseur J, Robben P. Learning from incidents in healthcare: the journey, not the arrival, matters. BMJ Qual Saf. 2017;26(3):252-6.
31. Vincent C, Amalberti R. Safety in healthcare is a moving target. BMJ Qual Saf. 2015;24(9):539-40.
32. Ovretveit J. The contribution of new social science research to patient safety. Soc Sci Med. 2009;69(12):1780-3.
33. Ovretveit J. Understanding and improving patient safety: the psychological, social and cultural dimensions. J Health Organ Manag. 2009;23(6):581-96.

Part III

New developments

Chapter 7

Detecting adverse events in clinical care using natural language processing

Klein DO, Rennenberg RJMW, Heuvel van den FAG, Koopmans RP, Prins MH

Submitted

Abstract

Objective: In this study, we wanted to find the best method (based on natural language processing (NLP)) to select cases out of the medical records for further investigation in search for a (potentially preventable) adverse event (AE).

Design: Retrospective medical record review compared with computer algorithm results.

Methods: The basic dataset consisted of 2987 medical records of patients who died during their hospitalization. To gain insight into the signal to noise ratio of the various resources, several subsets of our basic dataset were tested (first experiment). In the second experiment we tested the scalability. After the best subset was chosen, several NLP algorithms were tested to select the best performing algorithm for the detecting of AEs (third experiment). In the last experiment we tested the performance of the computer algorithms to predict potentially preventable AEs.

Results: The dataset with the last three letters showed the biggest potential. The scalability experiment showed that more data leads to a better performance of the algorithm. The best performing algorithm in the third test was the algorithm based on support vector machine (SVM), with a positive predictive value of 79%, a negative predictive value of 95% and a specificity of 85%. The results of the preventability experiment showed that the performance of the algorithms was almost equal to the results of the AEs.

Discussion and conclusion: In this study, we have shown that the SVM algorithm generates the most accurate results for the selection of cases for further investigation in the search for a (potentially preventable) AE. The sensitivity of the algorithms was around 75%. However, the SVM algorithm selected fewer cases to be examined for AEs compared to the original method. Consequently, this would lead to a lower workload for the committee. At the same time, there are a substantial number of cases, with potentially preventable AEs, not detected by machine learning.

Introduction

Current methods for retrospective review of medical records require both time- and cost-wise a substantial effort. Although trigger systems aim to decrease the burden, they still select many cases without adverse events (AEs). To optimise the process, sometimes cases are preselected leading to the investigation of a specific subset of cases. Examples are: deceased patients, patients of a particular department or patient with a longer length of stay. Results are therefore not generalizable to all hospital patients. Another issue is the relatively high inter-rater variation and low reproducibility of trigger results and subsequently AE detection with current methods.[1-7] This suggests that these methods generate results that make translation to clinical practice difficult.

To increase the efficacy, computer assisted detection of triggers, or even better AEs, in medical records has been developed. The aim is to reduce the labour-intensive manual chart review and to increase the reproducibility.[8] Thus far, text mining software is available for the detection of triggers or a limited selection of AEs.[9-17] The software has to be taught what triggers or AEs are, preferably with a large database of well characterized cases with their triggers and AEs. Until now, the most optimal dataset with patient details to be investigated is not known. For example, incorporating all information about the patients might influence the signal to noise ratio negatively and thus also the reliability of the results. Therefore, the optimal dataset has to be determined. Thereafter, different natural language processing (NLP) algorithms can be tested to detect the algorithm with the highest sensitivity and specificity for detecting triggers or AEs.

In this study, we wanted to find the most optimal method to select cases for further investigation in search for a (potentially preventable) AE. We describe our stepwise approach for developing a computer algorithm to search reliably for AEs in medical records. First, we used an NLP algorithm with excellent computing power restrictions to identify the optimal dataset selected from the medical record in our database with well characterized cases of deceased patients. Second, we evaluated the influence of the amount of records to be included for finding optimal results concerning agreement (true positives combined with true negatives divided by the total group). Third, several experiments were performed with different NLP algorithms in this optimal dataset to find the algorithm with the best performance for finding AEs and their preventability. Then, the performance of the best computer algorithm was validated in a different part of the data.

Methods

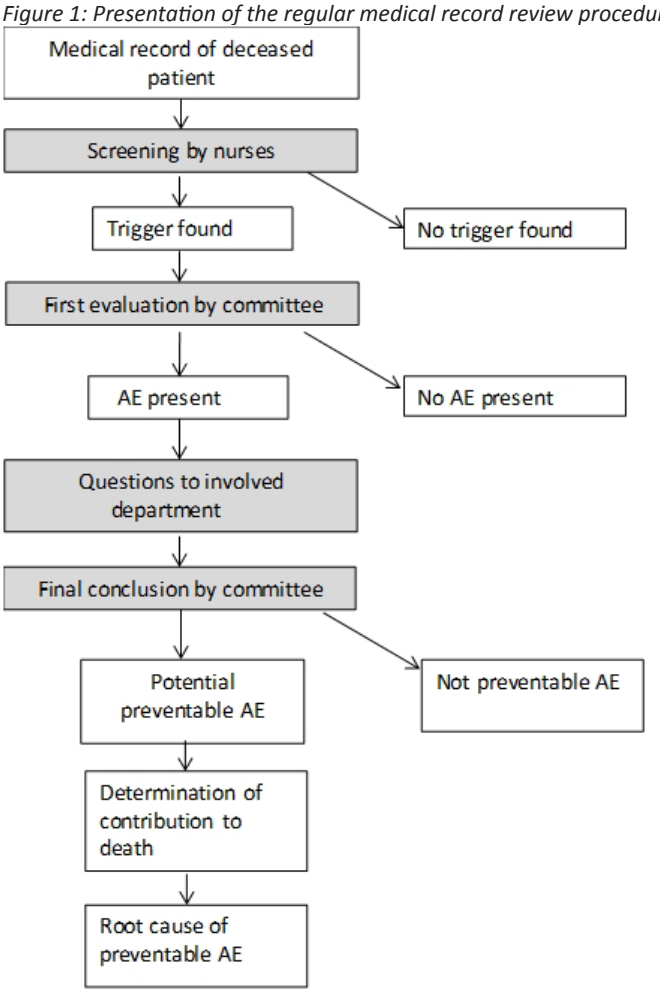
The usual procedure of the medical record review (our "gold standard")

Since 2008, a team of trained (according to the EMGO/NIVEL standards)[18] nurses screened the medical records of all deceased inpatients (approximately 700 annually) for the presence of triggers (figure 1). To accommodate the process, a database facilitating the necessary steps in this procedure was introduced in 2010 (Medirede®, Clinical File Search; Mediround BV). We used the triggers originally proposed by the Harvard medical practice study (HMPS) in 1991[19] with a slight adjustment to fit the group of deceased patients. Therefore, the triggers regarding transfer to another acute care hospital and unplanned inap-

propriate discharge to home were omitted as they have no relevance in deceased patients. The medical records with at least one trigger were redirected to the review committee. This committee consisted of both active and retired medical specialists with considerable clinical experience in the field of quality and safety in healthcare and medical record review. After a thorough review by a member specialised in the field of medicine related to the main diagnosis of the case (e.g. a surgeon investigates surgical patients etc.), this case was presented to the other members of the review committee in a regular meeting. A first conclusion on the potential presence of an AE was then established. Subsequently, after consulting the involved specialists, the committee finally decided on the presence of an AE and the potential preventability of this AE. Since 2012 there is a stable formation of the review committee.

Previous research showed that the average time nurses needed for the manual screening of triggers was 38 minutes (they had no time restrictions), for the reviewers this was on average 60 minutes (excluding the time needed for the discussion in a meeting). Thus, it takes approximately 1.5 hours for the total review of a single medical record.

Figure 1: Presentation of the regular medical record review procedure



Data

The basic dataset (originated from our "gold standard" procedure) consisted of 2987 medical records of patients who died during their hospitalization. All records between 2011 and 2016 were included. 1763 of these records (59%) contained one or more triggers after the screening by the nurses. In 742 of these medical records (42%) with a trigger, one or more AEs were detected by the review committee. 208 of these AEs were classified as potentially preventable. For these records, there was full access to surgical reports, discharge letters, patient records, nursing reports, use of medication, radiology reports, lab results and the medical history.

Definitions

An AE was defined as an unintended outcome arising from the (non)-action of a caregiver and/or the health care system with damage to the patient resulting in temporary or permanent disability or death of the patient.[20]
The clinical notes were defined as the document which describes the day-to-day report of the patient during admission. The patient file was defined as the document including all reports, letters, lab results, scans, reports and medical history.

Modified data

From the basic dataset several parts of data were selected. Machine learning is based on NLP. This is the ability of a computer program to understand human language as it is spoken. It is a component of artificial intelligence. In machine learning, it is important to make a selection out of data with a high signal-to-noise ratio. Preferably you would like to find the "signal" in the data, rather than fitting the noise. The signal was in this case the useful information in the medical record that is pointing towards the adverse event and the noise is the information in the medical record that is not helpful in locating/finding the adverse event. To gain insight into the signal to noise ratio of the various resources, several subsets of our basic dataset were tested (A-F).

A: Last general GP letter
In this selection, all records without a last general GP letter were excluded. There were 476 medical records without this letter, leaving 2511 (84%) for inclusion in the analysis. We have chosen the last GP letter because we assumed this contained the most useful information regarding the hospitalization of the patient.
B: Last letter
In this selection, the last general GP letter was used, but for the 476 cases in which this letter was missing, the last written document was used instead. Therefore, in this analysis, the original 2987 records were included.
C: Last three letters
In this analysis, the last three letters of all medical records were included.
D: Patient record
For this selection, the full patient record was used.
E: General GP letter combined with patient record
In this selection, the general GP letter was combined with the patient record, for every record with a general GP letter.
F: Last (general GP) letter combined with the edited patient record.
For this dataset, the patient record was electronically edited, leaving only 20% of the rarest words in the patient record. After the editing, the patient record was combined with dataset B. This was executed by a pre-processing script. First the whole text was evaluated and then the rare words were filtered out. After the editing, the patient record was combined with the last letter (dataset B).

Outcome measures

The following outcome measures were determined;
- Accuracy (agreement): calculated as the sum of true positives and true negatives

divided by the total population.
- Positive predictive value (PPV): calculated as the sum of true positives divided by the number of predicted condition positives (in this case AE present) = precision.
- Sensitivity (recall): calculated as the number of true positives divided by the total number of medical records which were identified as containing an AE.
- Negative predictive value (NPV): calculated as the number of true negatives divi ded by the number of predicted condition negative (in this case no AE present).
- Specificity: calculated as the number of true negatives divided by the total num ber of medical records which were identified as not containing an AE.

Computer NLP algorithms

We tested different computer algorithms and explored the feasibility of this software (Open Mines Platform supplied by "the Praktijk Index'').
The following NLP algorithms have been used:
- Naive Bayes (NB) with n-gram input and term frequency–inverse document fre quency (tf-idf) scores;
- Fast-text (FT) 2-layer neural network with hierarchical softmax output
- Linear Support Vector Machine (SVM)
- Convolutional neural networks (CNN) based on pre-trained word vectors[21,22]

Experiments

As a first step, the 6 datasets (described in section datasets) were provided to the NB algorithm to select the dataset that provided the highest performance in predicting an AE. This selected dataset was then used for the next experiments. To correct for the variation of the initialization, this experiment was repeated 28 times. Due to time and computing power restrictions, we chose to test the fast NB algorithm for all selections. In the second experiment (scalability) was tested whether the performance would decrease if a smaller training set was available. In the second experiment, we attempted to predict the preventability of an AE. AEs were therefore categorized as ''probably not preventable'' or ''potentially preventable''.
In the third experiment all four algorithms were trained, optimised and tested for AE prediction with use of the best-fitted dataset, which resulted from the first experiment (see box 1).

Box 1: Explanation of training, test and optimisation of the algorithm

1. Training set: this dataset consisted of examples used for the NLP to learn. In this case the NLP was taught which cases contained an AE and which didn't. learning, that is to fit the parameters (e.g., weights) of, for example, a classifier.[7][8]
2. Validation set: a set of examples used to tune the parameters of a classifier. It is sometimes also called the development set or the "dev set".
3. Test set : A test dataset is a dataset that is independent of the training dataset, but that fol lows the same probability distribution as the training dataset. A test set is therefore a set of examples used only to assess the performance (i.e. generalization) of a fully specified classi fier.[7][8]

60% of the data was used for training, 20% for validation and 20% for testing.

Privacy

To guarantee patients' privacy, information such as names, addresses, and other personal information were deleted from the data. After the anonymization, the data was checked by the privacy officer and researcher DK.

Results

Experiment 1: Dataset selection

In table 1 the results are presented for the individual datasets. The positive predictive value and the sensitivity varied widely for the different datasets. The accuracy was for all datasets around 75% and the specificity was close to 100% for every dataset.

Dataset C, which contained the last three letters, showed the biggest potential and was therefore used in the next experiment to test the performance of the different algorithms.

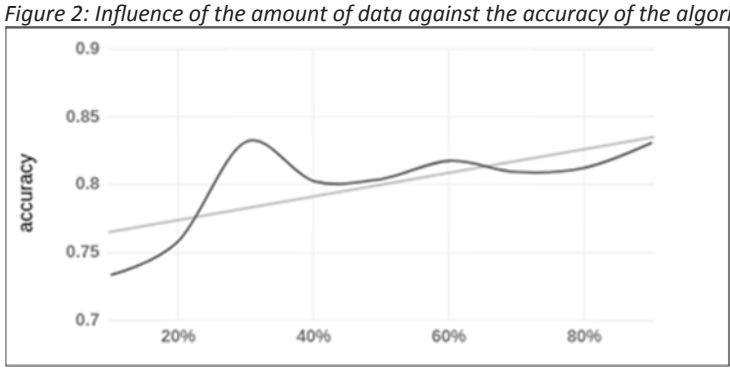
Table 1: Performance of the six datasets

Dataset	Accuracy	Precision (PPV)	Sensitivity (recall)	Specificity
A	74.9	66.8	75.1	98.8
B	75.3	72.1	75.2	99.2
C	76.1	78.3	76.1	99.7
D	75.8	57.0	76.0	100
E	75.7	57.0	76.0	100
F	74.9	67.6	74.9	99.0

Experiment 2: Scalability

Figure 2 below shows the influence of the amount of data plotted against the accuracy. This is shown for the NB algorithm. In this figure can be seen that more data leads to a better performance of the algorithm.

Figure 2: Influence of the amount of data against the accuracy of the algorithm



The size of the dataset is plotted against the accuracy of the NB algorithm. The resulting line is flattened and shown with corresponding trend line.

Experiment 3: Algorithm performances with dataset C

For accuracy, sensitivity and specificity the results were close together for the four algorithms (shown in table 2).

The best performing algorithm was the SVM algorithm, with a PPV of 75.8%, an accuracy of 82.3% and a specificity of 94.9%. Figure 3 shows the precision presented against the recall/sensitivity for the SVM algorithm. This shows if we would create a set with 82% precision, this would result in a recall/sensitivity of 40% (percentage AEs with respect to all AEs). In figure 4, the ROC curve of the SVM algorithm is presented. Table 2 shows the performance of the four algorithms separately. These results are based on 12 repetitions of the experiment, with the same settings for the algorithm but with a different distribution for the training set and the test set.

Figure 3: SVM algorithm with corresponding precision and accuracy

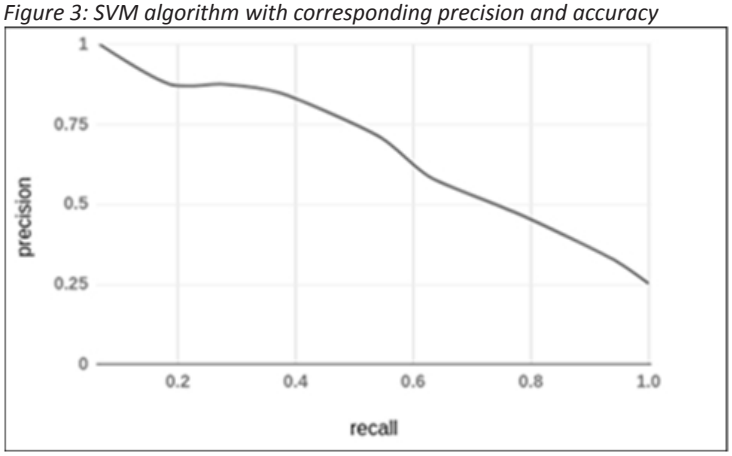


Figure 4: ROC curve of the SVM algorithm

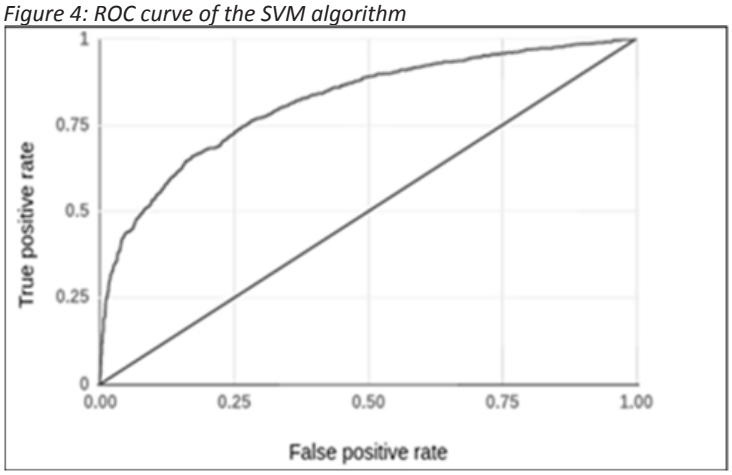


Table 2: Performance for every algorithm separately for dataset C

Algorithm	Accuracy	PPV	NPV	Sensitivity	Specificity
NB	0.76	0.67	0.77	0.11	0.98
SVM	0.84	0.79	0.95	0.51	0.85
FT	0.80	0.61	0.85	0.53	0.89
CNN	0.78	0.78	0.79	0.99	0.15

Experiment 4: Preventability

Table 3: Performance for the four algorithms separately for dataset C

Algorithm	Accuracy	PPV	Sensitivity	Specificity
NB	0.762	0.568	0.075	0.99
SVM	0.822	0.755	0.428	0.955
FT	0.798	0.610	0.513	0.89
CNN	0.778	0.730	0.173	0.98

The results of this experiment show that the performance of the algorithms for the preventability is almost equal to the results of the AEs. However, it must be considered that these results are based on a small dataset. This experiment was repeated five times.

Discussion and conclusion

In this study, we have shown that the SVM algorithm generates the most accurate results for the selection of cases for further investigation in the search for a (potentially preventable) AE. The sensitivity of the algorithms was around 75%. However, the SVM algorithm selected about 50% fewer cases to be examined for AEs compared to the original method. Consequently, this would lead to a lower workload for the committee.7 At the same time, there are a substantial number of cases, with potentially preventable AEs, not detected by machine learning. If we look at the best accuracy, then the SVM algorithm performs best. However, for precision, also the CNN algorithm performs well. The CNN and the SVM algorithm used the letters and the patient record. By using more structured information, which is already generated by hospitals, such as age and length of stay, the algorithms could be improved. Another option is using more data. The number of examples of which an algorithm will learn has an influence on the performance of the algorithm.
Because of an unfavorable signal to noise ratio in the total record the performance of the algorithms is dependent on the selected dataset. When different dataset selections are compared, it was indeed shown that the use of the unedited total patient record doesn't improve the output and can even have a negative effect on the results. Different datasets show different results for sensitivity and specificity. This is probably caused by the length of a patient record in itself and especially by long sentences.

The use of discharge letters, in particular, the letters for the GP, appeared to be useful because of a more favorable signal to noise ratio. These letters contain brief information regarding the hospitalization of the patient in contrast to the extensive patient record which contains a lot of "noise". The scalability experiment shows that the size of the dataset has

an effect on the precision of detecting AEs.

Finally, the results of the preventability experiment showed that the preventability can be adequately estimated despite scarce records in the training set.
Until now, the literature concerning the automatic detection of AEs focused on a short list of pre-specified AEs or only the triggers for which the data were searched. [9,11,14,16,17] Forster et al [23] and Murff et al (2003) screened discharge letters in two small patient samples for the presence of terms that could indicate an AE but the actual evaluation was performed by physician reviewers.[23,24] Murff showed a sensitivity of 69%, a specificity of 48% and a PPV of 52%. For Forster, the results were a sensitivity of 23%, specificity of 92%, PPV of 41% and NPV of 83%. In general, our results suggest a more optimal performance compared to these studies.

The strengths of our study were the careful step by step analysis with NLP to find the best algorithm that improved the PPV for a potentially preventable AE compared with the current trigger system. Several, currently common NLP algorithms were tested. Experiments were repeated until saturation was reached. Also, we tested different datasets to find the most useful one with the most valuable information.

Still, some questions remain because of some weaknesses in our study. First, it is not clear what an optimal cut off point for sensitivity or specificity could be. This has to be determined by further research to see if the data generated with this study make policy-makers draw the same conclusions as with conventional methods (in this study our "gold standard"). Second, the number of well-characterized records can be considered as rather small for NLP programs to be optimised, especially when we take the potential preventability into account. Finally, this was a selection of deceased patients and it might be even more interesting to use this method in patients that were alive at discharge from the hospital. Therefore, we think these data are not generalizable to living patients.

We think if the algorithms' performance is considered acceptable it should be tested prospectively and compared to the conventional methods. Considering that the conventional method is our "gold standard" the agreement will never be 100%. However, this is not an issue if the conclusions that can be drawn from the final results are comparable to the original ones. Still, we think the committee work cannot be fully replaced because the communication with the involved departments is crucial in the success of this quality and safety instrument.

In conclusion, the result of NLP algorithms to predict potentially preventable AEs in specific datasets from patient records is a promising tool to simplify record review. Further research is necessary, to investigate if the results of this method lead to the same overall conclusions from medical record review compared to the more expensive and labor-intensive conventional methods.

References

1. Sharek PJ, Parry G, Goldmann D, Bones K, Hackbarth A, Resar R, et al. Performance characteristics of a methodology to quantify adverse events over time in hospitalized patients. Health Serv Res. 2011;46(2):654-78.
2. Mattsson TO, Knudsen JL, Lauritsen J, Brixen K, Herrstedt J. Assessment of the global trigger tool to measure, monitor and evaluate patient safety in cancer patients: reliability concerns are raised. BMJ Qual Saf. 2013;22(7):571-9.
3. Hanskamp-Sebregts M, Zegers M, Vincent C, van Gurp PJ, de Vet HC, Wollersheim H. Measurement of patient safety: a systematic review of the reliability and validity of adverse event detection with record review. BMJ Open. 2016;6(8):e011078.
4. Zegers M, de Bruijne MC, Wagner C, Groenewegen PP, van der Wal G, de Vet HC. The inter-rater agreement of retrospective assessments of adverse events does not improve with two reviewers per patient record. J Clin Epidemiol. 2010;63(1):94-102.
5. Baines RJ, Langelaan M, de Bruijne MC, Wagner C. Is researching adverse events in hospital deaths a good way to describe patient safety in hospitals: a retrospective patient record review study. Bmj Open. 2015;5(7).
6. Klein DO, Rennenberg R, Koopmans RP, Prins MH. Adverse event detection by medical record review is reproducible, but the assessment of their preventability is not. PLoS One. 2018;13(11):e0208087.
7. Klein DO, Rennenberg R, Koopmans RP, Prins MH. The ability of triggers to retrospectively predict potentially preventable adverse events in a sample of deceased patients. Prev Med Rep. 2017;8:250-5.
8. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008:128-44.
9. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. Journal of the American Medical Informatics Association : JAMIA. 2005;12(4):448-57.
10. Penz JF, Wilcox AB, Hurdle JF. Automated identification of adverse events related to central venous catheters. Journal of biomedical informatics. 2007;40(2):174-82.
11. Rochefort CM, Buckeridge DL, Forster AJ. Accuracy of using automated methods for detecting adverse events from electronic health record data: a research protocol. Implementation science : IS. 2015;10:5.
12. Rochefort CM, Buckeridge DL, Tanguay A, Biron A, D'Aragon F, Wang S, et al. Accuracy and generalizability of using automated methods for identifying adverse events from electronic health record data: a validation study protocol. BMC Health Serv Res. 2017;17(1):147.
13. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA. 2011;306(8):848-55.
14. Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting adverse events using information technology. J Am Med Inform Assoc. 2003;10(2):115-28.
15. Stockwell DC, Kirkendall E, Muething SE, Kloppenborg E, Vinodrao H, Jacobs BR. Automated adverse event detection collaborative: electronic adverse event identification, classification, and corrective actions across academic pediatric institutions. Journal of patient safety. 2013;9(4):203-10.
16. Gerdes LU, Hardahl C. Text mining electronic health records to identify hospital adverse events. Studies in health technology and informatics. 2013;192:1145.
17. Sammer C, Miller S, Jones C, Nelson A, Garrett P, Classen D, et al. Developing and Evaluating an Automated All-Cause Harm Trigger System. Jt Comm J Qual Patient Saf. 2017;43(4):155-65.
18. Zegers M, de Bruijne MC, Wagner C, Groenewegen PP, Waaijman R, van der Wal G. Design of a retrospective patient record study on the occurrence of adverse events among patients in Dutch hospitals. BMC Health Serv Res. 2007;7:27.
19. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. N Engl J Med. 1991;324(6):370-6.
20. Wagner C. Onbedoelde schade in ziekenhuizen: resultaten dossieronderzoek naar patiëntveiligheid. Klachtenmanagement in de Zorg. 2007;4(3-4):28-31.
21. J H. Data Mining: Concepts and Techniques: Elsevier Science & Technology; 2011.
22. CM B. Pattern Recognition and Machine Learning: Springer-Verlag New York Inc.; 2006.
23. Forster AJ, Andrade J, Van Walraven C. Validation of a discharge summary term search method to detect adverse events. Journal of the American Medical Informatics Association. 2005;12(2):200-6.
24. Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. Journal of biomedical informatics. 2003;36(1-2):131-43

Chapter 8

Letter to the editor

Klein DO, Rennenberg RJMW

Preventive Medicine Reports; 2019

Letter to the Editor: A response to 'The ability of triggers to predict potentially preventable adverse events in a sample of deceased patients'

The article titled "The ability of triggers to predict potentially preventable adverse events in a sample of deceased patients"[1], published in the November 2017 issue of Preventive Medicine Reports, found a positive predictive value (PPV) of a trigger system in deceased patients to be 47%. Thereafter, we tried to optimise this PPV by adding characteristics (urgent admission, admission specialism) to the equation. However, this resulted only in a slightly better performance and the trigger method remained labour-intensive. Further research to optimise this system concerning the combination of triggers with patient characteristics and lab values seemed warranted. Additionally, literature showed that International Normalized Ratio (INR)[2], albumin-[3], creatinine-[4] and haemoglobin (Hb) levels[5] could be indicators for adverse events (AEs). Based on this information we decided to analyse our data with these variables. Instead of INR we included the use of anticoagulants and the number of different anticoagulants.

For this report, we extended the dataset to include a total of 4438 medical records of deceased patients (2011-2018). We created six algorithms using logistic regression with backward stepwise elimination to evaluate which variables contributed significantly to the prediction of the presence of AEs. Lab values were entered in the algorithm once as being measured yes or no, and once categorized as unmeasured, measured but normal, measured and abnormal using cut-off values used in the laboratory of our centre.

Table 1 shows the area under the curve (AUC) for these algorithms and sensitivity and specificity for a cut-off point chosen to optimise the algorithm's sensitivity. The AUC for the best predictive algorithm (algorithm 2) showed a value of 0.66 (with a p-value of <0.001). Adding lab results to the equation did not improve this and even the use of anticoagulants had no influence on the prediction. Therefore, we think adding lab results or use of anticoagulants to the equation does not improve the selection of cases with an AE.

Table 1: Algorithms for the prediction of the presence of AEs

Algorithms	AUC	95% CI	P-value	Sensitivity	Specificity	Cut-point
1: based on age and gender	0.52	0.503-0.541	0.022	0.60	0.46	0.27
2: 1 + length of stay	0.66	0.646-0.682	<0.001	0.70	0.54	0.23
3: 2+ measurement of lab values	0.66	0.646-0.682	<0.001	0.70	0.54	0.23
4: 2 + albumin, Hb, glucose, creatinine, CRP	0.66	0.644-0.680	<0.001	0.70	0.53	0.23
5: 4 + use of anticoagulants and number of different anticoagulants.	0.67	0.648-0.684	<0.001	0.70	0.53	0.23
6: 3+ use of anticoagulants and number of different anticoagulants.	0.67	0.649-0.685	<0.001	0.70	0.53	0.23

References

1. Klein DO, Rennenberg R, Koopmans RP, Prins MH. The ability of triggers to retrospectively predict potentially preventable adverse events in a sample of deceased patients. Prev Med Rep. 2017;8:250-5.
2. Oden A, Fahlen M. Oral anticoagulation and risk of death: a medical record linkage study. BMJ. 2002;325(7372):1073-5.
3. Seo SH, Kim SE, Kang YK, Ryoo BY, Ryu MH, Jeong JH, et al. Association of nutritional status-related indices and chemotherapy-induced adverse events in gastric cancer patients. BMC Cancer. 2016;16(1):900.
4. Santopinto JJ, Fox KA, Goldberg RJ, Budaj A, Pinero G, Avezum A, et al. Creatinine clearance and adverse hospital outcomes in patients with acute coronary syndromes: findings from the global registry of acute coronary events (GRACE). Heart. 2003;89(9):1003-8.
5. Ammann RA, Niggli FK, Leibundgut K, Teuffel O, Bodmer N. Exploring the association of hemoglobin level and adverse events in children with cancer presenting with fever in neutropenia. PLoS One. 2014;9(7):e101696.

General discussion

Discussion

The general aim of this thesis was to evaluate the performance of medical record review (MRR) as an instrument to measure and improve patient safety in clinical care. This method consisted of the Harvard Medical Practice Study (HMPS) trigger tool, although slightly adapted and only applied to patients who died during their stay in our hospital.[1]
We calculated the predictive value of the triggers (chapter 3) and the reproducibility of the triggers (chapter 4). Furthermore, we assessed the inter- and intrarater reproducibility for the decision on the presence of an adverse event (AE), the potential preventability and the contribution to death of a patient (chapter 5). After that, we assessed the external reproducibility of MRR regarding this second stage (chapter 6).
Additionally, we investigated if an automatic trigger tool could be helpful in the selection of cases with a (potentially preventable) AE (chapter 7). Finally, we evaluated the potential improvement of the trigger system by adding lab values (chapter 8).
In this chapter, we will provide an overview of our main findings also in relation to existing literature. Moreover, we will touch upon methodological considerations of our studies and future recommendations.

Main findings

Although MRR is performed on a large scale (both for study purposes as well as in daily practice) in health care institutions throughout the world, we concluded from this thesis that this method needs some improvement to make it more (cost)effective. Hereafter we will first summarize the highlights of this thesis and then discuss them.

Review

In our review of the literature we found that MRR is a reasonably well studied method for the evaluation of medical records for AEs.[2-15] Still, looking at the WHO quality criteria for screening instruments, much research concerning the separate quality requirements is still lacking or the methodology is questionable. Especially research concerning the cost of detecting AEs, valuable information is missing regarding the cost of the method itself.16 Moreover, knowledge on how MRR changes quality and safety for patients has hardly been investigated and, to our opinion, should be evaluated.

Triggers

We demonstrated that the current trigger system has a positive predictive value (PPV) of 47%, which results in a selection of medical records without an AE in more than 50% of the cases.[17] However, these results are slightly better compared to other studies, with an average PPV of 33% for the HMPS triggers.[2-5,9,12,14,18-20] Unfortunately, a comparison of the PPV of the individual triggers isn't possible since these have not been reported on an individual level in previous studies.

Although this trigger system seems to perform better in our subset of patients we still think the PPV is disappointingly low. However, as we concluded from this study, in our opinion not all triggers are equally reliable for selecting records with a potential AE. This

was shown by a (on average) moderate kappa for the individual triggers (with a range between 0 and 0.78).[21] Nonetheless, our results for the observed agreement of the individual triggers was higher than in other studies. Unbeck et al (2014) reported a reproducibility of 46% compared to a 67% reproducibility in our study. In addition, the total agreement for any trigger present was 65% in their study, compared to 90% in our study.[22] Again, no comparison was possible for the individual triggers in the studies presented in the literature. Our results therefore seem slightly better. However, this might be due to the fact that we only use records of deceased patients which generally contain the most serious events and perhaps therefore most easy detectable triggers.[23] Moreover, our trained nurses are very experienced in searching for triggers in medical records because several years of screening and training preceded this study.

Improvement of the predictive value of the trigger system

After the evaluation of the predictive value of the trigger system we also attempted to improve this. As a first step we added patient characteristics to the trigger system. This resulted in a lower number of records which were selected for review than in the original review (PPV is 50% compared to 47% previously).
As a second step, we added laboratory results to the trigger system to further improve this PPV. The PPV was now 54% in the algorithm with the best area under the curve. Which is unfortunately still a disappointing figure but also gives a distorted image since also a large percentage of AEs is missed.
Ideally, the trigger system would have a PPV of nearly 100%. This would minimize the review of medical records without AEs and also lower the cost of the whole process. However, this would probably mean that AEs would be missed in records that are not selected by the system. A weakness of this current study is that we were not aware of potential AEs in records that were previously not selected for extensive review due to the lack of a trigger.[17]

AE

We were convinced that using internal reviewers combined with a group discussion for the determination of an AE and the use of a 3 level scale for the determination of preventability, would lead to a higher internal reliability compared to other studies in which external reviewers or a pair of reviewers were used.[4,24-30] Our study showed a substantial agreement for the presence of an AE. However, for the potential preventability the kappa had only a fair agreement. When we calculated the intrarater and interrater agreement concerning potential preventability, the kappa showed that, mainly for interrater agreement, the reproducibility was low.[31]
The result of our external reproducibility study showed a moderate kappa for the presence of an AE. For preventability of an AE the agreement was fair and it was poor for contribution of an AE to death.[32] We believe that this could be caused by lack of a clear definition for preventability.[33]
Preventability is a difficult concept. Although there is usually consensus about the extremes, there is a large grey area. To improve this agreement on preventability, we are convinced there is need for a more uniform definition. Although a six-point scale, like some other studies use, seems to show a higher reproducibility with a moderate kappa, we think

this is still not a satisfying result.[34] Moreover, a multilevel scale suggests precision that seems disputable because of the moderate reproducibility.[12]

MRR seems not useful for the evaluation of individual cases, as is shown by the disappointing results of the reproducibility. However, we still believe that MRR is suitable to detect trends in AEs. This might then give rise to management decisions aimed to diminish the occurrence of certain types of AEs. For example, the assignment of a thrombosis vigilance officer to lower the occurrence of improper use of anticoagulants in patients and hence AEs concerning bleeding or thrombosis.

Text mining

Text mining was developed as an instrument to quickly scan plain text with great accuracy. In our study where we explored the possibilities of automatic screening of medical records for triggers and AEs, we have shown that it is possible to select cases with a high probability of having an AE. A disadvantage of this method is that, when including more text documents of a single medical record, the results became less precise. To address this problem, we selected specific documents from the database (such as the last letter to the general practitioner) that led to results with more precision. Compared to the selection of cases by trained nurses applying triggers, only a small part of the medical records were selected by the computer. This could mean that AEs are missed in records that are not selected for review by the text mining software. However, this does not have to be a problem if the general conclusion from the small amount of selected cases and their AEs, after thorough review, show the same trend in AEs and possible preventable causes. The final conclusions in the annual report from the committee should then also be the same. However, this wasn't investigated in this study and should be further explored in future studies.
Still, the screening possibilities for huge amounts of data by text mining tools are much greater than the screening possibilities for humans. Here, we selected cases of deceased patients only, but it could easily be used to scan medical records of all patients (in less time). The program is also likely to improve because it is self-learning. The more data is used, the better the prediction algorithm will become like in other artificial intelligence setups with big data. Concerning big data, our data volume is much too small to expect the program to improve the results revolutionary.
Another disadvantage of this study was that we compared the results of the automatic detection with the ''gold standard''. Our Gold Standard is the normal procedure with trained nurses. This means that the automatic selection could never show better results than the current manual evaluation of the medical records.

Methodological considerations

MRR is often subject of debate whether it is the best method for the evaluation of patient safety. The main arguments for this are listed below:

General

Marking your own paper
An often-heard argument against MRR by an internal review committee is that it is seen as ''marking your own paper''. But from experience with MRR in our centre, we can conclude that if you would let an external committee evaluate the medical records, they don't have the insights in the local health care process that an internal committee has. A view that is shared by others.[24] On the other hand, if one encourages specialists (of the involved department), instead of a separate committee, to evaluate their own medical records, chances are that almost no AEs are found. Therefore, we believe that it is best when medical records are evaluated by an internal committee in which representatives of several departments are present.[24]

Hindsight bias
Hindsight bias, also known as the knew-it-all-along phenomenon refers to the common tendency for people to perceive events that have already occurred as having been more predictable than they actually were before the events took place. It also plays a role in retrospective MRR, where the reviewers already know the outcome. In this view, it is difficult to judge decisions made with the available information at the time of the decision without knowing what is really going to happen. It is difficult for reviewers, policymakers and families to acknowledge that a correct decision was made that turned out to be false as time progressed. It is therefore easy to classify a decision as false, but one has to look at the dilemma with only the available information at the time of the dilemma. Questions regarding medical decisions when reviewing files of deceased patients are difficult to rise without prejudice induced by hindsight. They should be inquiring about why certain decisions were made and why other options were not pursued or considered. If sufficient effort was deployed to retrieve the necessary information and the reasoning with the available information was correct, there is no AE. Regardless of the fact that in the end (with more information in or after time progressed) this turned out to be wrong.[35-40]

Retrospective
Another disadvantage of a retrospective method is that it can be only reactive. Which means that AEs are evaluated after they occurred. In fact, you are always ''running behind''. An advantage of a prospective measuring method would be, that one would be able to prevent AEs.[41] However, this would necessitate information about predictors of an AE in individual patients which, to our knowledge, are not available yet. Moreover, it will be difficult to find predictors that are uniformly applicable to hospitals. It will probably be necessary for every hospital to develop these predictors specific for patients in their own hospital. This also will necessitate the analysis of big data to make them as precise as possible. Moreover, one should then still investigate if an intervention in the identified patient would indeed result in the prevention of an AE.

Specific for our studies

Sample size
Due to time and cost restrictions, the sample sizes of our internal and external reproducibility studies were small. Still, we based our sample size for the reproducibility of the AEs on the study of Walter et al.[42]
For the sample size calculation of the internal reproducibility we estimated that a sample size of 50 records would provide sufficient power to exclude a kappa of as low as 0.6 given an expected kappa of 0.8. For the external reproducibility a sample size of 80 records would provide sufficient statistical power to exclude a kappa of as low as 0.6 given an expected kappa of 0.75. However, the results from both these studies indicated that the true Kappa is much lower than anticipated. Hence, we were not able to reject the null-hypothesis that Kappa is not significantly different from 0.6.

Triggers
In the study where we evaluated the reproducibility of the triggers, more triggers were found the second time that the medical records were evaluated. We suspect that extra attention among the nurses due to the fact that the second round of review was part of a study might have contributed to this. Also, we found AEs in cases without triggers. However, the percentage of AEs in cases without triggers was much lower (7%) than in the cases with a trigger (47%).[21]

External reproducibility
For the external reproducibility study, the two committees were not entirely comparable. There was a difference in experience and also a difference in the composition of the committee. Additionally, although the committees were instructed to use their common method for review and final decision we cannot exclude any influence of the fact that the review of the 40 cases in the other hospital was done especially for study purposes. Finally, a substantial number of the medical records contained more than 1 AE, which made it easier for the external committee to find at least one of these AEs; this could have led to an overestimation of the external reproducibility.

Patient selection
We chose to evaluate the medical records of patients who died during their stay. In our opinion, the most severe cases of AEs can be found in patients that eventually die. Moreover, in deceased patients twice as much preventable AEs can be found. Therefore, investigation of records in this population appears to be more efficient for identifying preventable AEs than reviewing records of patients who were discharged alive.[23]

Generalizability

The results of our external reproducibility study showed that the lack of a clear definition on preventability is a major problem. We believe that this problem is generalizable to other large hospitals in the Netherlands. This makes comparison of preventability results of such committees between hospitals, a challenge if not impossible. A single external committee might prevent this to a certain level. However, as explained previously, we think that external committees are less valuable in helping to improve local procedures. Besides that, the use of a better definition concerning preventability might also improve the internal reproducibility in other hospitals. Although discussion between experts generates a general conclusion, the lack of a clear definition can lead to a different conclusion when the same case is discussed again on a different moment amongst the same experts. Furthermore, our results concerning text mining are only representative for our hospital. Although one could think this is generalizable to others, we think that this should be proven by repeating these investigations in other hospitals (external validation).

Future recommendations

This thesis has shown that, to our opinion, MRR needs some improvement. They are listed below.

Definitions

First, it is important that the definitions used in MRR are better specified. In particular the definition of the preventability of an AE. Several severity scales are used in the literature for preventability. For the outcome, some consist of only two options (yes/no) but there are also systems that make you choose for a graded scale outcome between 4, 5 or even 6 options.[4,24,27,43-46] It isn't known if the outcomes are related to the conditions. The latter suggests achieving precision in the assessment of preventability is possible. However, there is to our knowledge no sound research proving this precision to be reproducible.[12,34,47] We are convinced that there is a need for a clear definition that can be widely used and accepted. This could be done on a national level in the Netherlands, although it would be even better to have an international standardisation. In that case, hospitals in different countries could also compare their outcomes. A Delphi round amongst experts seems the best solution to define these criteria for preventability in the future. In the end, the classification of an AE is now highly dependent on the skills, experience and beliefs of the medical record investigator.

Yield

Second, the percentage of records that is triggered is relatively high with eventually a low yield of AEs.[17] Hence, it would be beneficial if we could devise a system in which the actual percentage of records that contain a preventable AE after evaluation is higher. This could be achieved by optimizing our selection instruments or the selection of the population in which we look for AEs. A promising future option is the deployment of computer algorithms designed after big data analysis.[48-53] A computer is able to screen an enormous number of variables and measure their influence on adverse outcomes. Hence, it might even be possible to predict a chance for a patient to suffer from an AE in the near future. This would enable us to design and test preventive strategies. Until now, the reviewer's decision is considered the "gold standard". But is this the best we can get? Because there are so many variables it might be impossible for humans to balance them repeatedly in the same manner. Therefore, computer algorithms might be better equipped to manage big data. Still, the acceptance of computer output that can only be explained with difficulty will probably be very low. Computer results always have to be translated to workable plans that can be grasped by people on the work floor. Hence, the reasons for this classification by a computer should always be as clear as possible.

Costs

Third, an evaluation of the costs of MRR versus the effect on the daily practice of the healthcare workers and thereafter the results on AEs should be carefully evaluated. However, there are to our knowledge no well validated instruments that measure the complete cycle (detection, interpretation, feedback, change in care, outcome). We are of opinion that such programs or instruments should be thoroughly evaluated for their exact costs and

benefits. We then could incorporate them in daily practice but only if the costs of such a screening program are acceptable in terms of cost per quality adjusted life year (QALY).[54-56] This might be complicated because although some improvements in safety seem obvious, changes in daily practice could cause "collateral" damage. For example, double check of intravenously administered medication by nurses seems a good idea.[57] But the time spent on the double check cannot be spent on other tasks. Which choices will then be made and will this lead to possible other unexpected AEs for patients.[58,59] Moreover, we should realise that improving the quality and safety of an already high-quality system means an exponential increase in costs.[60] That obligates us to perform thorough research on this matter and finally incorporate the most optimal strategies in our health care.[61]

Improvement of care

This thesis investigated the current MRR process in our centre and possibilities for improvement. It would also be of interest to expand the research and evaluate if and how this process affects the improvement of care. The whole process could then be evaluated. This is also shown in the overview figure 1 in the introduction.

Conclusion

In short, medical record review can be used to evaluate the trend of adverse events in a hospital. Due to its low reproducibility it is less useful for the evaluation of individual cases. To improve the reproducibility of preventability, it is advisable to define a uniform definition for preventability. Also, an evaluation of the costs versus the effect is necessary.

References

1. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. N Engl J Med. 1991;324(6):370-6.
2. Vincent C, Neale G, Woloshynowych M. Adverse events in British hospitals: preliminary retrospective record review. BMJ. 2001;322(7285):517-9.
3. Davis P, Lay-Yee R, Briant R, Ali W, Scott A, Schug S. Adverse events in New Zealand public hospitals I: occurrence and impact. N Z Med J. 2002;115(1167):U271.
4. Baker GR, Norton PG, Flintoft V, Blais R, Brown A, Cox J, et al. The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2004;170(11):1678-86.
5. Thomas EJ, Studdert DM, Burstin HR, Orav EJ, Zeena T, Williams EJ, et al. Incidence and types of adverse events and negligent care in Utah and Colorado. Medical care. 2000;38(3):261-71.
6. Wilson RM, Runciman WB, Gibberd RW, Harrison BT, Newby L, Hamilton JD. The Quality in Australian Health Care Study. The Medical journal of Australia. 1995;163(9):458-71.
7. Ock M, Lee SI, Jo MW, Lee JY, Kim SH. Assessing reliability of medical record reviews for the detection of hospital adverse events. Journal of Preventive Medicine and Public Health. 2015;48(5):239-48.
8. de Feijter JM, de Grave WS, Muijtjens AM, Scherpbier AJ, Koopmans RP. A comprehensive overview of medical error in hospitals using incident-reporting systems, patient complaints and chart review of inpatient deaths. PloS one. 2012;7(2):e31125.
9. Sari AB, Sheldon TA, Cracknell A, Turnbull A. Sensitivity of routine system for reporting patient safety incidents in an NHS hospital: retrospective patient case note review. BMJ (Clinical research ed). 2007;334(7584):79.
10. Macharia WM, Muteshi CM, Wanyonyi SZ, Mukaindo AM, Ismail A, Ekea H, et al. Comparison of the prevalence and characteristics of inpatient adverse events using medical records review and incident reporting. S Afr Med J. 2016;106(10):1021-36.
11. Unbeck M, Schildmeijer K, Henriksson P, Jurgensen U, Muren O, Nilsson L, et al. Is detection of adverse events affected by record review methodology? An evaluation of the " Harvard Medical Practice Study" method and the " Global Trigger Tool". Patient safety in surgery. 2013;7 (1) (no pagination)(10).
12. Zegers M, de Bruijne MC, Wagner C, Hoonhout LH, Waaijman R, Smits M, et al. Adverse events and potentially preventable deaths in Dutch hospitals: results of a retrospective patient record review study. Qual Saf Health Care. 2009;18(4):297-302.
13. Hanskamp-Sebregts M, Zegers M, Vincent C, Van Gurp PJ, Vet HCWD, Wollersheim H. Measurement of patient safety: A systematic review of the reliability and validity of adverse event detection with record review. BMJ open. 2016;6 (8) (no pagination)(e011078).
14. Sousa P, Uva AS, Serranheira F, Nunes C, Leite ES. Estimating the incidence of adverse events in Portuguese hospitals: a contribution to improving quality and patient safety. BMC health services research. 2014;14:311.
15. Kobayashi M, Ikeda S, Kitazawa N, Sakai H. Validity of retrospective review of medical records as a means of identifying adverse events: Comparison between medical records and accident reports. Journal of evaluation in clinical practice. 2008;14(1):126-30.
16. Bates DW, O'Neil AC, Petersen LA, Lee TH, Brennan TA. Evaluation of screening criteria for adverse events in medical patients. Medical care. 1995;33(5):452-62.
17. Klein DO, Rennenberg R, Koopmans RP, Prins MH. The ability of triggers to retrospectively predict potentially preventable adverse events in a sample of deceased patients. Prev Med Rep. 2017;8:250-5.
18. Unbeck M, Schildmeijer K, Henriksson P, Jurgensen U, Muren O, Nilsson L, et al. Is detection of adverse events affected by record review methodology? an evaluation of the "Harvard Medical Practice Study" method and the "Global Trigger Tool". Patient Saf Surg. 2013;7(1):10.
19. Akbari Sari A, Doshmangir L, Torabi F, Rashidian A, Sedaghat M, Ghomi R, et al. The Incidence, Nature and Consequences of Adverse Events in Iranian Hospitals. Arch Iran Med. 2015;18(12):811-5.
20. Soop M, Fryksmark U, Koster M, Haglund B. The incidence of adverse events in Swedish hospitals: a retrospective medical record review study. International journal for quality in health care : journal of the International Society for Quality in Health Care. 2009;21(4):285-91.
21. Klein DO, Rennenberg R, Koopmans RP, Prins MH. The Harvard medical practice study trigger system performance in deceased patients. BMC Health Serv Res. 2019;19(1):16.
22. Unbeck M, Lindemalm S, Nydert P, Ygge BM, Nylen U, Berglund C, et al. Validation of triggers and development of a pediatric trigger tool to identify adverse events. BMC Health Serv Res. 2014;14:655.
23. Baines RJ, Langelaan M, de Bruijne MC, Wagner C. Is researching adverse events in hospital deaths a good way to describe patient safety in hospitals: a retrospective patient record review study. BMJ Open. 2015;5(7):e007380.
24. Sharek PJ, Parry G, Goldmann D, Bones K, Hackbarth A, Resar R, et al. Performance characteristics of a metho-

dology to quantify adverse events over time in hospitalized patients. Health Serv Res. 2011;46(2):654-78.
25. Zegers M, de Bruijne MC, Wagner C, Groenewegen PP, Waaijman R, van der Wal G. Design of a retrospective patient record study on the occurrence of adverse events among patients in Dutch hospitals. BMC Health Serv Res. 2007;7:27.
26. Asavaroengchai S, Sriratanaban J, Hiransuthikul N, Supachutikul A. Identifying adverse events in hospitalized patients using Global Trigger Tool in Thailand. Asian Biomedicine. 2009;3(5):545-50.
27. Landrigan CP, Parry GJ, Bones CB, Hackbarth AD, Goldmann DA, Sharek PJ. Temporal trends in rates of patient harm resulting from medical care. The New England journal of medicine. 2010;363(22):2124-34.
28. Kurutkan MN, Usta E, Orhan F, Simsekler MC. Application of the IHI Global Trigger Tool in measuring the adverse event rate in a Turkish healthcare setting. The International journal of risk & safety in medicine. 2015;27(1):11-21.
29. Classen DC, Resar R, Griffin F, Federico F, Frankel T, Kimmel N, et al. 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. Health affairs (Project Hope). 2011;30(4):581-9.
30. Hwang JI, Chin HJ, Chang YS. Characteristics associated with the occurrence of adverse events: A retrospective medical record review using the Global Trigger Tool in a fully digitalized tertiary teaching hospital in Korea. Journal of evaluation in clinical practice. 2014;20(1):27-35.
31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.
32. Klein DO, Rennenberg RJMW, Gans ROB, Enting RH, Koopmans RP, Prins MH. Limited external reproducibility restricts the use of medical record review for benchmarking. BMJ Open quality 2019;8.
33. Nabhan M, Elraiyah T, Brown DR, Dilling J, LeBlanc A, Montori VM, et al. What is preventable harm in healthcare? A systematic review of definitions. BMC health services research. 2012;12:128.
34. Klein DO, Rennenberg R, Koopmans RP, Prins MH. Adverse event detection by medical record review is reproducible, but the assessment of their preventability is not. PLoS One. 2018;13(11):e0208087.
35. Fischhoff B. Hindsight not equal to foresight: the effect of outcome knowledge on judgment under uncertainty. 1975. Qual Saf Health Care. 2003;12(4):304-11; discussion 11-2.
36. Henriksen K, Kaplan H. Hindsight bias, outcome knowledge and adaptive learning. Qual Saf Health Care. 2003;12 Suppl 2:ii46-50.
37. Hugh TB, Dekker SW. Hindsight bias and outcome bias in the social construction of medical negligence: a review. J Law Med. 2009;16(5):846-57.
38. Arkes HR. The Consequences of the Hindsight Bias in Medical Decision Making. Curr Dir Psychol Sci. 2013;22(5):356-60.
39. Banham-Hall E, Stevens S. Hindsight bias critically impacts on clinicians' assessment of care quality in retrospective case note review. Clin Med (Lond). 2019;19(1):16-21.
40. B. F. Hindsight not equal to foresight: the effect of outcome knowledge on judgment under uncertainty. . Exp Psychol: Human Percept Perform. 1975;1:288-99.
41. Michel P, Quenon JL, de Sarasqueta AM, Scemama O. Comparison of three methods for estimating rates of adverse events and rates of preventable adverse events in acute care hospitals. BMJ. 2004;328(7433):199.
42. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med. 1998;17(1):101-10.
43. Hayward RA, Hofer TP. Estimating hospital deaths due to medical errors: preventability is in the eye of the reviewer. JAMA. 2001;286(4):415-20.
44. Schumock GT, Thornton JP. Focusing on the preventability of adverse drug reactions. Hospital pharmacy. 1992;27(6):538.
45. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, et al. Incidence of adverse events and negligence in hospitalized patients: results of the Harvard Medical Practice Study I. 1991. Quality & safety in health care. 2004;13(2):145-51; discussion 51-52.
46. Nilsson L, Risberg MB, Montgomery A, Sjodahl R, Schildmeijer K, Rutberg H. Preventable Adverse Events in Surgical Care in Sweden: A Nationwide Review of Patient Notes. Medicine. 2016;95(11):e3047.
47. Aranaz-Andres JM, Aibar-Remon C, Limon-Ramirez R, Amarilla A, Restrepo FR, Urroz O, et al. Prevalence of adverse events in the hospitals of five Latin American countries: results of the 'Iberoamerican Study of Adverse Events' (IBEAS). BMJ Qual Saf. 2011;20(12):1043-51.
48. Rochefort CM, Buckeridge DL, Forster AJ. Accuracy of using automated methods for detecting adverse events from electronic health record data: a research protocol. Implementation science : IS. 2015;10:5.
49. Caron F, Vanthienen J, Vanhaecht K, Van Limbergen E, Deweerdt J, Baesens B. A process mining-based investigation of adverse events in care processes. The HIM journal. 2014;43(1):16-25.
50. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. Journal of the American Medical Informatics Association : JAMIA. 2005;12(4):448-57.

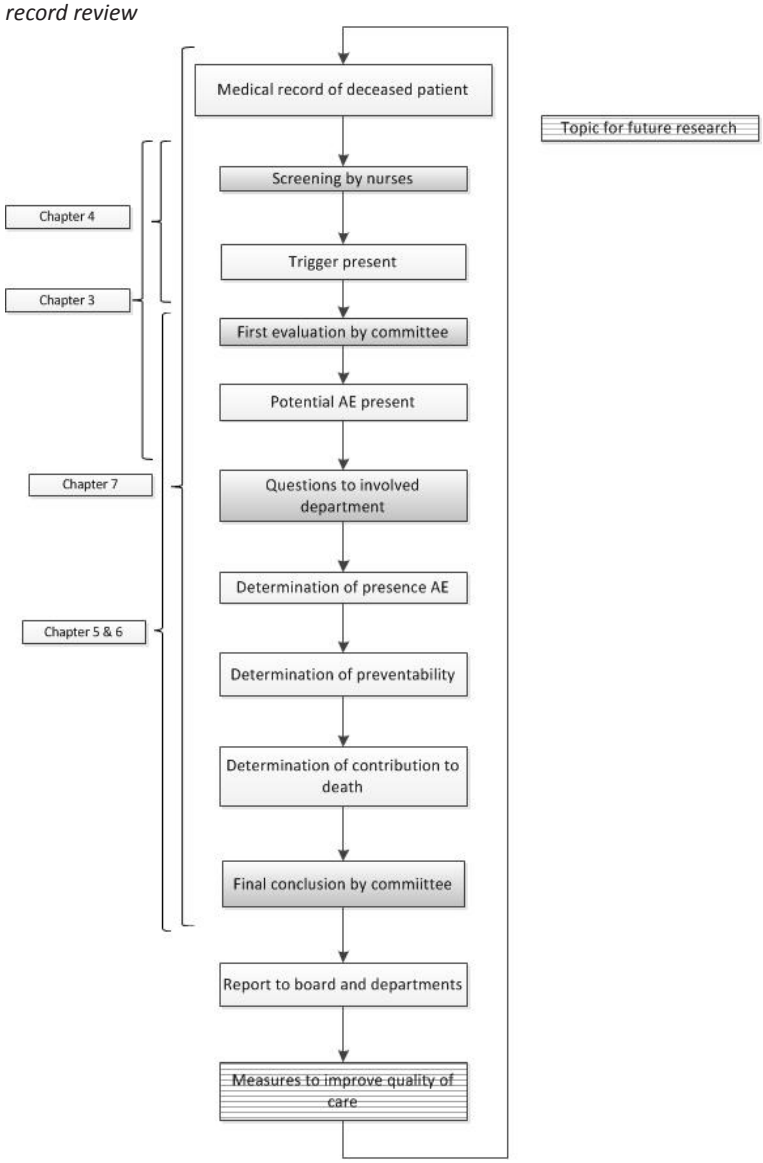
51. Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting adverse events using information technology. Journal of the American Medical Informatics Association : JAMIA. 2003;10(2):115-28.
52. Stockwell DC, Kirkendall E, Muething SE, Kloppenborg E, Vinodrao H, Jacobs BR. Automated adverse event detection collaborative: electronic adverse event identification, classification, and corrective actions across academic pediatric institutions. Journal of patient safety. 2013;9(4):203-10.
53. Schwendimann R, Blatter C, Dhaini S, Simon M, Ausserhofer D. The occurrence, types, consequences and preventability of in-hospital adverse events - a scoping review. BMC health services research. 2018;18(1):521.
54. Carter AW, Mandavia R, Mayer E, Marti J, Mossialos E, Darzi A. Systematic review of economic analyses in patient safety: a protocol designed to measure development in the scope and quality of evidence. BMJ Open. 2017;7(8):e017089.
55. Zsifkovits J, Zuba, M., Geissler, W., Lepuschütz, L., Pertl, D., Kernstock, E., Osterman,, H. . Costs of unsafe care and cost effectiveness of patient safety programmes. European Commissien 2016.
56. Ziektelast in de praktijk - De theorie en praktijk van het berekenen van ziektelast bij pakketbeoordelingen. Zorginstituut Nederland 2018.
57. Schwappach DL, Pfeiffer Y, Taxis K. Medication double-checking procedures in clinical practice: a cross-sectional survey of oncology nurses' experiences. BMJ Open. 2016;6(6):e011394.
58. Hewitt T, Chreim S, Forster A. Double checking: a second look. J Eval Clin Pract. 2016;22(2):267-74.
59. Alsulami Z, Conroy S, Choonara I. Double checking the administration of medicines: what is the evidence? A systematic review. Arch Dis Child. 2012;97(9):833-7.
60. Juran JM. Classical algorithm of optimum quality costs. McGraw-Hill; 1988. p. From Jurans Quality Control Handbook, 4th edition.
61. Øvretveit J. Does improving quality save money? The Health Foundation 2009.

Summary

Summary

In this thesis we investigated how reproducible the medical record review of deceased patients is for the estimation of the quality of the delivered care. For this research we highlighted different aspects of medical record review and we subdivided this process in its various sub processes.
In the introduction (chapter 1) we describe how quality and safety is evaluated in hospitals. Furthermore, we explain the method of medical record review, which is used in Maastricht UMC+. Also, we describe the outline of this thesis (figure 1).

Figure 1: Overview of our studies regarding the process of improving quality and safety by medical record review



Part I

In part I we evaluated what has been investigated in the domain of quality and safety in health care. The review article describes the literature specifically for two trigger tools which can be deployed for medical record review (chapter 2).

For this review we evaluated the literature based on the 7 steps which have been composed by the WHO for the evaluation of methods.

50 studies met the inclusion criteria and we discussed the results stepwise. We found that with medical record review, more AEs are detected than with any other method. But at the same time, other AEs are found. Each method detects some AEs that are not detected with other methods.
The average costs of an AE were €4296.

Considerable efforts have been made worldwide in the healthcare systems to improve safety and to reduce errors in the treatment of patients. These efforts have resulted in some positive effects.

The literature also showed that medical record review focused on several domains of quality of care such as safety and effectiveness and seems suitable for both small as large cohorts.

Furthermore, we found a moderate agreement for the presence of a trigger and a good agreement for the presence of an AE for the GTT across studies. For the HMPS we found a substantial agreement for the presence of a trigger and a moderate agreement for the presence of an AE.

Medical record review, regardless of the trigger tool used, is a reasonably well researched method for the evaluation of the medical records for AEs. However, looking at the WHO criteria, much research is still lacking or of moderate quality. Especially for the cost of AEs, valuable information is missing regarding the costs of their detection. Moreover, knowledge of how MRR changes quality and safety of care should be evaluated.

Part II

In chapter 3 we investigated the performance of the triggers regarding the prediction of the presence of a (potentially preventable) AE. Furthermore, we evaluated the possibility to increase this predictive value. Our study showed that the total predictive value was 47%. If more triggers were present in a case, this would increase the chance of detecting an AE.

By adaptation of the trigger system (for example: adding length of stay, age, gender and admission from another hospital) it seemed possible to increase the predictive value considerably. But, we should keep in mind that a part of the AE could be missed (depending on the chosen cut-off value).

In chapter 4 we evaluated the reproducibility of the 15 individual triggers. For this study, the nurses evaluated 100 medical records for a second time. The observed agreement for the presence of a trigger was 75% with a moderate kappa. We found that certain triggers such as unplanned transfer to the ICU and unplanned readmission to the operating room had a higher reproducibility than others such as dissatisfaction with care and adverse drug event. For the individual triggers the agreement was 90% and the corresponding kappa moderate.

In chapter 5 we investigated the internal reproducibility of the committee outcome. Therefore, the committee evaluated their previously processed medical records for a second time. This was done for presence of an AE, the potential preventability and the contribution to death. Also, we again evaluated the root cause of the AEs. Kappa for the presence of an AE had a substantial agreement and a fair agreement for the (potential) preventability. The intrarater agreement showed a substantial agreement for the presence of an AE and for a (potential) preventable AE. The interrater agreement showed a substantial agreement for the AE presence and a slight agreement for a (potential) preventable AE. We concluded that an international consensus on a definition for preventability is needed. This would results in a better comparison between studies.

In chapter 6 we evaluated the external reproducibility of the committee outcome. Therefore, a second committee reviewed previously reviewed medical record, and vice versa. The two committees functioned as each other's external committee. From both hospitals we selected 40 medical records, which had been checked by the internal committee in the previous years. After the second evaluation round we calculated the total agreement and the kappa agreement. We did this for the presence of an AE, the potential preventability and the contribution to death. Kappa for the presence of AEs was moderate. The kappa for preventability was weak and poor for the contribution to death.

Although both committees applied the same methods for medical record review, there is a difference in results, possibly because the committees differ in their experience with MRR (3 vs. 8 years) and have different medical backgrounds. However, we think that medical record is suitable for the detection of general issues regarding patient safety and also for the discussion of individual cases. However, the suboptimal reproducibility of MRR reduces its potential for benchmarking. Finally, we think at least a better definition of preventability and also of contribution to death is needed if we want to compare the outcomes between hospitals.

Part III

In chapter 7 we investigated the usefulness of an automatic trigger tool for the screening of medical records for triggers and AEs. For this study we compared the results of medical records which had been evaluated by the committee in the previous years (2011-2016) with the outcome of different computer algorithms.

We defined various datasets as derived from the electronic patient record as input, to check which dataset would perform best. After this we investigated the scalability. In the third experiment we tested four different natural language processing (NLP) algorithms to

find the algorithm which performed best in the detection of AEs. In the last experiment we investigated if it would be possible to use NLP in the search for preventable AEs. The subset which contained the last three letters of the medical records had the biggest potential and was therefore used for the next experiments. The scalability experiment showed that more data leads to a better performance of the computer algorithm. From the third experiment we concluded that support vector machine (SVM) leads to the best results with a positive predictive value of 79% with a negative predictive value of 95% and a specificity of 85%.

The results of the last experiment were comparable with this. The SVM algorithm selected fewer cases with a (potential) AE. This would lead to a lower workload of the committee. But at the same time a considerably amount of the AE wouldn't be detected with this new method.

Chapter 8 describes the (potential) improvement of the predictive value of the current trigger tool, by adding laboratory results and other patient characteristics. This study elaborates on the research described in chapter three. In the analyses we added the following variables: albumin, creatinine, haemoglobin, glucose, use of anticoagulants and number of anticoagulants. Our total dataset contained 4438 medical records of deceased patients (2011-2018). We created six algorithms based on backward logistic regression, to evaluate which combination of variables had the best predictive values.

The area under the curve for the best predictive algorithm showed a value of 0.73 (with a p-value of <0.001), which means that a patient with an AE will have a more abnormal test result than 70% of the patients without an AE. Although in this algorithm only 3 variables (age, gender and length of stay) are included, further inclusion of more variables didn't further improve the predictive value. Based on these results it seems that adding laboratory results doesn't improve the PPV of the trigger system considerably.

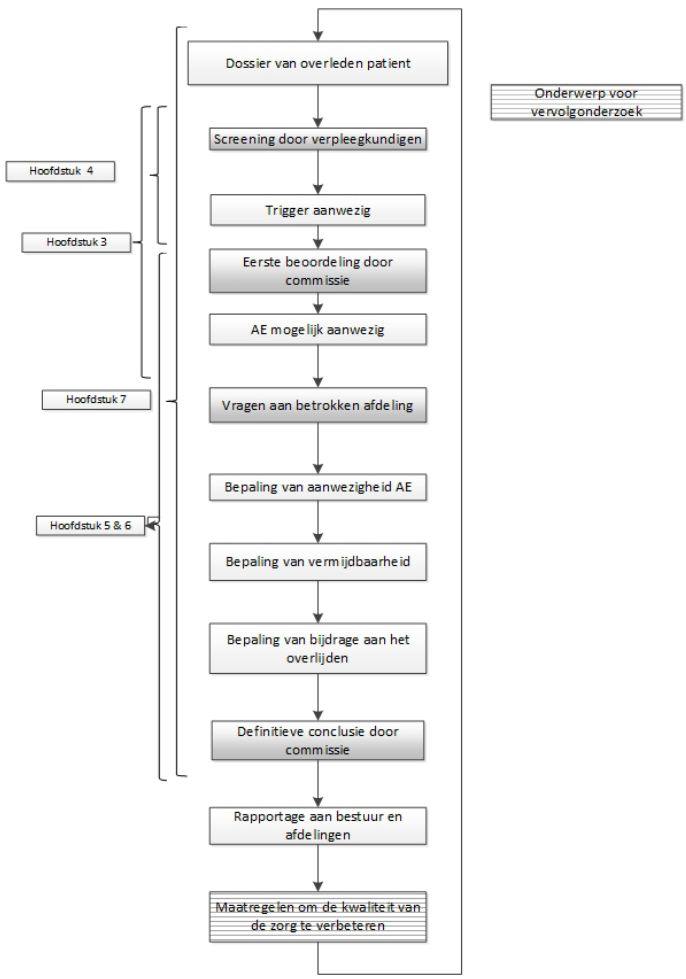
Samenvatting

Samenvatting

Om de kwaliteit van zorg te meten beschikken zorginstellingen over diverse meetinstrumenten. Een daarvan is dossieronderzoek. Dit proefschrift behelst een onderzoek naar de kwaliteit en reproduceerbaarheid van dossieronderzoek van overleden patiënten. Het betreft alleen overleden patiënten omdat de gedachte is dat in die groep de meest ernstige, en daardoor beter zichtbare, onbedoelde schade gevonden kan worden. Voor dit onderzoek hebben we de verschillende kanten belicht van dat dossieronderzoek en het hele proces van selectie en analyse van dossiers in stappen verdeeld.

In de introductie (hoofdstuk 1) omschrijven we de hoe kwaliteit en veiligheid geëvalueerd wordt in ziekenhuizen. Daarnaast wordt de methode van dossieronderzoek, zoals deze wordt toegepast in het Maastricht UMC+, uitgelegd. Ook wordt de opzet van dit promotieonderzoek besproken, dit wordt ook getoond in afbeelding 1.

Figuur 1: Overzicht van onze studies betreffende het proces van de verbetering van kwaliteit en veiligheid door middel van dossieronderzoek



Deel I

In deel I hebben we literatuur gezocht om te achterhalen wat er al bekend is over de kwaliteit en de reproduceerbaarheid van dossieronderzoek zoals deze gebruikt wordt voor het meten van kwaliteit en veiligheid in de gezondheidszorg. Vervolgens hebben we een uitgebreide analyse gedaan van deze literatuur waarvan we de resultaten beschrijven in hoofdstuk 2. Dit artikel beschrijft de literatuur specifiek op het gebied van twee instrumenten die gebruikt worden om aanwijzingen voor onbedoelde schade op te sporen (trigger tool) en die ingezet worden voor dossieronderzoek van patiënten. De literatuur werd geanalyseerd met behulp van zeven criteria die door de WHO zijn opgesteld voor het evalueren van screeningsmethoden.

50 studies voldeden aan onze inclusiecriteria en de resultaten werden stapsgewijs besproken. Zo vonden we dat er met behulp van dossieronderzoek meer AEs gevonden wordt dan met andere methoden. Maar tegelijkertijd leveren verschillende methoden andere AEs op. De gemiddelde kosten van een enkel geval van onbedoelde schade waren €4296. Wereldwijd zijn er aanzienlijke inspanningen gedaan om de patiëntveiligheid te verbeteren en fouten te verminderen. Dit heeft in sommige gevallen geleid tot een kleine daling in gevallen van onbedoelde schade.

Met dossieronderzoek wordt ingegaan op verschillende domeinen van kwaliteit van de zorg zoals veiligheid en effectiviteit, daarnaast kan het ingezet worden om zowel kleine als grote groepen te onderzoeken. Met dossieronderzoek wordt een matige tot goede overeenstemming voor de aanwezigheid van een aanwijzing van onbedoelde schade en de aanwezigheid van werkelijk onbedoelde schade gevonden, afhankelijk van het gebruikte instrument.

Uit de resultaten van dit onderzoek hebben we kunnen afleiden dat er de afgelopen decennia redelijk veel onderzoek gedaan is naar dossieronderzoek. Maar, met de WHO criteria in het achterhoofd is de kwaliteit van deze onderzoeken matig. Daarnaast wordt er in onderzoek wel aandacht besteedt aan de kosten van onbedoelde schade zelf, maar niet aan de kosten van het detecteren van onbedoelde schade. Die laatsten zijn aanzienlijk en van alle kwaliteitsinstrumenten samen in Nederland in 2015 al 80 miljoen euro per jaar. Om die reden is het ook van groot belang dat er meer kennis wordt verworven over hoe de veranderingen in kwaliteit en veiligheid het beste kunnen worden geëvalueerd.

Deel II

In hoofdstuk 3 hebben we onderzocht hoe goed vijftien verschillende aanwijzingen voor onbedoelde schade kunnen voorspellen dat een dossier een (potentieel) vermijdbare onbedoelde schade bevat. Verder hebben we geanalyseerd of het mogelijk was om deze voorspellende waarde te verhogen. Uit ons onderzoek bleek de totale voorspellende waarde 47% te zijn. Hoe meer aanwijzingen aanwezig waren in een dossier, hoe groter de kans was op het vinden van onbedoelde schade. Door aanpassingen aan het systeem (o.a. door toevoeging van opnameduur, herkomst, geslacht en leeftijd) was het mogelijk om de positieve voorspellende waarde aanzienlijk te verhogen. Wel moet er dan rekening gehouden worden met het feit dat een deel van de onbedoelde schade gemist zal worden

(afhankelijk van het gekozen afkappunt).

In hoofdstuk 4 is de reproduceerbaarheid van de 15 individuele aanwijzingen voor onbedoelde schade onderzocht. Hiervoor hebben de verpleegkundigen 100 dossiers voor een tweede keer beoordeeld. De geobserveerde overstemming voor de aanwezigheid van een aanwijzing voor onbedoelde schade was 75% met een kappa van 0.5 (kappa is een voor kans gecorrigeerde maat van overeenkomst tussen beoordelingen)102. Uit onderzoek bleek dat bepaalde aanwijzingen zoals ongeplande overplaatsing naar de intensive care en ongeplande heroperatie een hogere reproduceerbaarheid hadden dan anderen zoals ontevredenheid met de zorg en bijwerking van de medicatie. Voor de individuele aanwijzingen was de overeenstemming 90% en de bijbehorende kappa 0.42 (range: -0.03-0.78).

In hoofdstuk 5 hebben we de interne reproduceerbaarheid van het oordeel van de commissie onderzocht, door de commissie een tweede keer hun eigen dossiers te laten bekijken. Dit is gedaan op het gebied van de aanwezigheid van een onbedoelde schade, de vermijdbaarheid en de bijdrage aan het overlijden van deze gebeurtenis. Ook hebben we gekeken naar de oorzaak van de onbedoelde schades. De kappa voor de aanwezigheid van een onbedoelde schade was substantieel en redelijk voor de (potentiele) vermijdbaarheid. De overeenstemming tussen beoordelaars had een substantiële kappa voor de aanwezigheid van een onbedoelde schade en voor een (potentieel) vermijdbare onbedoelde schade. De overeenstemming met een eigen eerdere oordeel van de beoordelaars had ook een substantiële kappa van en een geringe kappa voor een (potentieel) vermijdbare onbedoelde schade.

In hoofdstuk 6 hebben we onderzocht wat de externe reproduceerbaarheid is van het commissie oordeel als twee commissies elkaars dossiers bekijken. De twee commissies fungeerden dus als elkaars externe commissie. Van beiden ziekenhuizen werden 40 dossiers geselecteerd die in de voorafgaande jaren door de interne commissie beoordeeld was. Na het tweede oordeel door de andere (dus externe) commissie, hebben we de totale overeenstemming en de kappa berekend. Dit hebben we gedaan voor de aanwezigheid van een onbedoelde schade, de vermijdbaarheid en de bijdrage aan het overlijden.
De kappa voor de aanwezigheid van onbedoelde schades was matig. Voor de vermijdbaarheid was dit zwak en zelfs slecht voor de bijdrage aan het overlijden. Ondanks dat beide commissies dezelfde methode voor het dossieronderzoek hanteerde is er een verschil in de resultaten te zien. Deze verschillen kunnen te verklaren zijn door de verschillen in ervaring van de commissie, of door de verschillende medische achtergrond van de commissieleden. Desondanks zijn we nog steeds van mening dat dossieronderzoek een geschikte methode is om discussiepunten te detecteren aangaande patiëntveiligheid. Deze kunnen dan intern besproken worden en eventueel leiden tot een veranderde aanpak van een probleem waarmee de patiëntveiligheid verbeterd. Echter, op basis van de resultaten van dit onderzoek bevelen we aan om voorzichtig om te gaan met het gebruik van dossieronderzoek als referentiekader (benchmarking). Ten slotte zijn we van mening dat een betere definitie van vermijdbaarheid en bijdrage aan het overlijden nodig zijn om vergelijkingen tussen ziekenhuizen te kunnen maken.

Deel III

In hoofdstuk 7 hebben we onderzocht of het mogelijk was om automatisch in dossiers naar aanwijzingen voor onbedoelde schade te screenen en naar daadwerkelijke onbedoelde schade.

Voor dit onderzoek hebben we de resultaten van 2987 dossiers die in de voorgaande jaren (2011-2016) geëvalueerd waren door de commissie, vergeleken met de uitkomsten van verschillende computeralgoritmes. Om na te gaan welke dataset het beste zou functioneren, hebben we in het eerste experiment diverse selecties samengesteld. Vervolgens hebben we de schaalbaarheid onderzocht. Nadat de beste selectie gekozen was, hebben we vier verschillende algoritmes getest die werken op basis van natuurlijke taalverwerking (natural language processing) om het algoritme te vinden die als beste onbedoelde schade kan detecteren (experiment drie). In het laatste experiment hebben we onderzocht of het mogelijk is om met behulp van NLP ook vermijdbare onbedoelde schade te vinden.
De selectie met daarin de laatste drie brieven van het dossiers, had de grootste potentie en werden daarom gebruikt voor de andere experimenten. Het experiment waarin de schaalbaarheid werd onderzocht, toonde aan dat meer data leidt tot een betere prestatie van het computeralgoritme. Uit het derde experiment bleek dat het support vector machine (SVM) algoritme leidde tot de beste resultaten met een positief voorspellende waarde van 79%, een negatief voorspellende waarde van 95% en een specificiteit van 85%. De resultaten van het laatste experiment waren hiermee vergelijkbaar. Het SVM algoritme selecteerde minder casus met daarin mogelijk een onbedoelde schade. Dit zou leiden tot een lagere belasting voor de commissie. Maar tegelijkertijd zal een aanzienlijk deel van de casus niet gevonden worden met behulp van deze nieuwe methode.

Hoofdstuk 8 beschrijft de (potentiele) verbetering van de voorspellende waarde van het huidige instrument gebruikt voor het vinden van aanwijzingen voor onbedoelde schade, door het toevoegen van laboratoriumwaarden en andere patiëntkenmerken. Dit onderzoek borduurt voort op het onderzoek dat beschreven is in hoofdstuk drie. We hebben in de analyses buiten de variabelen die we al eerder hadden toegevoegd aan de aanwijzingen (leeftijd, geslacht, opnameduur, opnamespecialisme, spoedopname, aanwezigheid ontslagbrief, opname vanuit een ander ziekenhuis en aantal aanwijzingen) de volgende toegevoegd: albumine, creatinine, hemoglobine, INR (maat voor stolbaarheid van het bloed), gebruik van antistollingsmedicatie, het aantal verschillende antistollingsmedicatie en het percentage gewichtsverlies.

Voor de analyses gebruikten we de gegevens van 4438 dossiers waarin een of meer van deze variabelen bekend was. Uit de logistische regressie bleek dat er een aantal variabelen significant bij te dragen aan de kans om onbedoelde schade te ondervinden. Dit zijn; leeftijd, geslacht, spoedopname, albumine, glucose, hemoglobine, creatinine, gebruik van antistolling en het aantal verschillende antistollingsmedicijnen. We hebben 6 algorithmlen samengesteld gebaseerd op ''backward'' logistische regressie om te bepalen welke variabele de beste voorspellende waarde heeft. Het algorithm met daarin leeftijd, geslacht en opnameduur had de beste voorspellende waarde. Verdere toevoeging van labwaarden zorgde niet voor een aanzienlijke verbetering van de voorspellende waarde.

Valorisation

Valorisation

Target audience

The results of this thesis could be of relevance for a wide audience, namely hospitals in general, policymakers, quality officers, physicians, researchers and most of all for patients. However, we think the results should mainly be used by hospital quality officers who have to implement preventive strategies for improvement in quality and safety control. Moreover, it stresses the necessity to investigate other instruments for their effectiveness.

Social relevance

Nobody wants to be a victim of unintended medical harm. To prevent this, it is necessary to identify risky situations. The methods we use to identify these situations should be precise and valid because otherwise we might change clinical practice in a way that won't solve the problem. Preventing unintended harm to patients is the key role of these instruments and methods to detect adverse events (AEs) in health care. Therefore, it is of utmost importance that these methods are of good quality. In this thesis we aimed to improve our knowledge about the precision and performance of a commonly used method for detecting AEs. This knowledge enables us to increase the accuracy of the method and therefore optimise the chance for better choices in improving health care safety. Better instruments with well-known characteristics are expected to result in more benefits to society. First by less harm inflicted by health care and second by lower costs.

Economic relevance

In the Netherlands, 10% of patients who die during their stay in hospital experience an AE according to the medical record review (MRR) method.[1] According to Hoogervorst-Schilp et al (2015), the costs of AEs in the Netherlands are around €300 million each year.[2] A substantial number of these AEs was considered preventable and these costs could therefore be reduced.

These extra costs are caused by the impact on the patient itself and its family, and also by additional - and costly- care. Moreover, there are substantial extra costs for the patient, insurance companies and also for the economy since every extra day a patient stays in the hospital leads to inevitable 'costs' through lost income, payments by insurers etc. Therefore, it is necessary to use a test with optimal test characteristics to detect these AEs and use the results to prevent them in the future. Moreover, scarce research has been performed regarding the balance between the costs of these methods themselves and the savings from safer healthcare they are eventually meant realise. Therefore, we think that spending a substantial part of our healthcare budget on tests and instruments that give blurry answers does not seem like a good idea.

Relevance for patients

Improving the measurements of healthcare quality and patient safety gives us more reliable and precise information about the delivered care. It then enables us probably to target our efforts in making care safer with better quality. Ideally this would subsequently result in less AEs and less extra days in the hospital and is therefore of importance for patients. At the moment, medical record review is suitable for evaluating the trends of AEs. It is less suitable for individual cases, which has been shown in this thesis by its low reproducibility.

Relevance of measurements

By studying the properties of this commonly used method we learn more about the precision with which it tells us about medical situations that carry substantial risks for adverse events. If this information is reliable, we can advise professionals and policy makers to think about reducing risks in these situations. The better the quality of the measurements the better the advice and, in our opinion, the greater the chance that changes will really improve quality and safety. In this thesis we did not yet measure the specific results of MRR on changes in policy or care and finally the effect on the occurrence of AEs. Medical record review with the use of triggers (as shown in this thesis) gives valuable information on the potential presence of an AE, its potential preventability and the contribution to death.

Discussing the committees' findings with the involved health care providers might urge them to adapt protocols and workflow in order to improve quality and safety of the delivered care. Also, the right instrument is needed to measure the change in AEs. While there are other methods available for measuring patient safety and quality, medical record review is seen as the gold standard.

Relevance for health care providers

Based on the results in this study, we advise health care providers to evaluate detection and measuring tools before they implement them. If a decent instrument is used, this will not only give the best results but also will be cost- and timesaving compared to the use of an instrument which is not optimal. Also, the use of a method which has been shown to be a valid method is more justified than a method which hasn't been evaluated.

Innovation

Besides the evaluation of the current method, we have evaluated new options to improve of the detection tool for AEs. We added readily available variables to the trigger tool method and deleted less discriminative ones. We also explored the possibilities for artificial intelligence to help us with searching for AEs. This might make the detection process more efficient in the future.

Future perspectives

First and most important future perspective is an evaluation of the costs of MRR versus the effect on daily practice of health caretakers and thereafter the results on AEs should be done. However, this is a real challenge because care is developing quickly and it will be difficult to tell which improvements are the result of MRR indirectly or were realised by general improvements in care.

Second, harmonisation in the definitions used in MRR is better specified. In particular the term preventability.

Third, the percentage of records that are selected for the possible presence of an AE, with an actual AE after evaluation should be higher to make the time and costs consuming part of manually scrutinizing records more efficient.

Fourth, the investigation on automatic selection of MRR and AEs has only been briefly researched by us. More research is warranted because this might make it possible to investigate all records and not just a selection.

Fifth, the next step in the evaluation of the quality of MRR is the measurement of the effect of outcomes of the committee on the daily practice in the hospital. So, do the health caretakers adapt their way of working after the findings of the MRR and also does this adaptation results in less (preventable) AEs?

Finally, the evaluation of the MRR is a first analysis of one of the instruments that are used in the hospital but this evaluation can also be used for other instruments. Such as the central incident committee, complaint committee, complaint service point, quality parameters registries.

References

1. Langelaan M, Broekens, M.A., de Bruijne, M.C., de Groot, J.F., Moesker, M.J., Porte, P.J., Schutijzer, B., Singotani R., Smits, M., Zwaan, L., Asscherman, H., Wagner, C. . Monitor Zorggerelateerde Schade 2015/2016 - Dossieronderzoek bij overleden patiënten in Nederlandse ziekenhuizen. 2017.
2. Hoogervorst-Schilp J, Langelaan M, Spreeuwenberg P, de Bruijne MC, Wagner C. Excess length of stay and economic consequences of adverse events in Dutch hospital patients. BMC Health Serv Res. 2015;15:531.

Curriculum vitae

Dorthe Odyl Klein was born in Roermond, on the 14th of September in 1985.
After graduating from secondary school (Bisschoppelijk College Broekhin, Roermond) in 2004, she studied Nutrition & Dietetics at the Hogeschool van Arnhem en Nijmegen, in Nijmegen, which she completed in 2008. Subsequently she obtained the master Nutrition and Health, with specialisation Nutrition in Health and disease, at the University of Wageningen in 2010.
After this, she has worked for several years as a scientific knowledge manager at Danone Research in Wageningen. This was followed by a job as a research dietitian for the Encore study at the department of Epidemiology at the University of Maastricht. In March 2015 she joined the Department of Clinical epidemiology and health technology assessment (KEMTA) at Maastricht UMC+ to obtain her PhD degree. She was supervised by dr. Roger Rennenberg, prof. dr. Richard Koopmans and prof. dr. Martin Prins. This resulted in the work presented in this thesis.
At the moment she is still working at Maastricht UMC+, but now for the department of Quality and Safety.

List of publications

Published papers

Klein DO; Rennenberg RJMW: A response to 'The ability of triggers to predict potentially preventable adverse events in a sample of deceased patients'. Preventive Medicine Reports. 2019; 10.1016/j.pmedr.2019.100920.

Klein DO; Rennenberg RJMW; Gans ROB; Enting RH; Koopmans RP; Prins MH. Limited external reproducibility restricts the use of medical record review for benchmarking. BMJ Open Quality. 2019. 8: e000564.

Klein DO; Rennenberg RJMW, Koopmans RP, Prins MH. The Harvard medical practice study trigger system performance in deceased patients. BMC Health Services Research. 2019. 19(16).

Klein DO, Rennenberg RJMW, Koopmans RP, Prins MH. Adverse event detection by medical record review is reproducible, but the assessment of their preventability is not. PLoS One. 2018. 13(11): e0208087.

Klein DO, Rennenberg RJMW, Koopmans RP, Prins MH. The ability of triggers to retrospectively predict potentially preventable adverse events in a sample of deceased patients. Preventive Medicine Reports. 2017;8:250-255.

Submitted for publication

Klein DO, Rennenberg RJMW, Koopmans RP, Prins MH. A narrative systematic review of medical record analysis to improve patient safety in hospitals: is this method evidence based?

Klein DO, Rennenberg RJMW, van den Heuvel FAG, Koopmans RP, Prins MH. Detecting adverse events in clinical care using natural language processing

Presentations

Using artificial intelligence in healthcare for the detection of AEs.
Onderlinge tafel voor data scientists in de zorg. Oral presentation. MediRisk, May 2019.

Artificial intelligence and the detection of AEs.
Gebruikersdag Performation HOTflo. Oral presentation. Bilthoven, February 2019.

Does text mining software detect adverse events reliably?
International forum on quality and safety in healthcare. Oral presentation. Amsterdam, May 2018.

Retrospective medical record review as a quality instrument to measure patient safety. CaRe PhD day. Oral Presentation. Maastricht. May 2017.

Patient safety: characteristics related to adverse events.
International forum on quality and safety in healthcare. Poster presentation. Gotenburg, May 2016.

Design of the research for the evaluation of medical record review.
Bijeenkomst samenwerkende topklinische opleidingsziekenhuizen (STZ) Zuid. Oral presentation. Maastricht, March 2016.

Dankwoord

Zoals gebruikelijk zal ook ik dit proefschrift afsluiten met een woord van dank aan iedereen die mij de afgelopen jaren in meer of mindere mate heeft bijgestaan.

Allereerst wil ik beginnen met het bedanken van mijn promotieteam!

Roger, als copromotor heb je mij tijdens mijn promotie met veel enthousiasme begeleidt. Bedankt voor de vrijheid die je me gegeven hebt, hieruit blijkt je vertrouwen in mij en daar ben ik je heel dankbaar voor. Eindeloos veel tijd heb ik doorgebracht in de stafgang van niveau 5, maar ik heb ervan geleerd dat wachten loont! Geintje natuurlijk….Ik wil je heel erg bedanken voor de fijne samenwerking tijdens de afgelopen jaren. Ik hoop dat we nog vele onderzoeken samen mogen uitvoeren. Aan de ideeën die we hebben zal het niet liggen, daar kunnen we nog wel een aantal promotietrajecten mee vullen!

Daarnaast wil ik mijn promotoren, Richard Koopmans en Martin Prins bedanken. Richard, door jouw heldere kijk op de zaken zijn de artikelen alleen maar duidelijker geworden. Daarnaast ben ik je dankbaar dat je me in het eerste jaar meteen liet weten dat ik alvast moest bedenken wat ik hierna wilde gaan doen. Mede hierdoor ben ik me er altijd van bewust gebleven dat er nog een leven is na het promoveren! Ik ben daardoor ook op tijd begonnen met het uitzoeken wat ik wil gaan doen in mijn verdere loopbaan. Martin, veel van onze afspraken vonden plaats via Skype, een uitstekend medium om snel en efficiënt te werken. Bedankt voor het redigeren van de stukken en de statistische input!

Dit gehele onderzoek was niet mogelijk geweest zonder de steun van de Raad van Bestuur van het Maastricht UMC+. Hartelijk dank daarvoor!

Tevens hebben we gedurende het onderzoek diverse keren de (financiële) hulp van Hans Fiolet moeten inschakelen. Dank voor uw bijdrage! Daarnaast ben ik heel blij dat we elkaar nog even gesproken hebben voordat mijn aanstelling als promovendus ten einde was. Dankzij u kan ik nog even blijven binnen het Maastricht UMC+.

Beste (oud)leden van de COOP – Harry, Peter K, Peter S, Pierre, Chris, Mark, Paul Br, Roger, Arno, Jan S, Jan T, Jan V, Paul Be, Fabienne en Trang. Zonder jullie was dit hele onderzoek er niet geweest, de vele jaren van dataverzameling hebben ervoor gezorgd dat ik al bij de start van mijn onderzoek een schat aan informatie tot mijn beschikking had. Daarnaast nog een extra woord van dank voor Chris, Peter K en Pierre, die samen met mij naar Groningen zijn gereisd om daar in een kamer afgesloten van de buitenwereld dossiers te beoordelen voor het onderzoek naar de externe reproduceerbaarheid.

Beste (oud)leden van het triggerteam: Wil, Bianca, Wilma, Nicole, Iris, Guido, Karin – jullie zijn ook een onmisbaar onderdeel van mijn onderzoek geweest. Bedankt voor jullie inzet! Graag wil ik ook de leden van de leescommissie bedanken voor het lezen en beoordelen van dit proefschrift: Prof. Dr. C.A.B. Webers, Prof. Dr. C.G. Faber, Prof. Dr. S.E. Geerlings, Dr. L.J.G.G. Panis en Prof. Dr. C. Wagner.

Astrid, als secretaris van de COOP heb je me ontelbaar veel keren geholpen met van alles en nog wat. Heel erg bedankt!

Beste Sjef, zonder jou had ik nooit de juiste informatie gehad omtrent de opname- en ontslagtijd. Bedankt!

Beste Arno, ik vond het heel fijn om met jou samen te werken tijdens het project voor de Praktijk Index. Bedankt voor de tijd en de moeite die je erin gestoken hebt!

Beste André en beste Frans, bedankt voor de fijne samenwerking tijdens ons project over machine learning!

Wim en Cor, jullie wil ik natuurlijk ook bedanken voor al de gegevens die door de jaren heen verzameld zijn met behulp van Medirede en voor de ideeën die jullie nog in de loop der jaren hebben bedacht voor vervolgonderzoek. Ik hoop van harte dat het nieuwe project snel van start kan gaan!

Tijdens een van onze projecten hebben we samengewerkt met het UMCG. Daarvoor zijn drie leden van de COOP aldaar, drie dagen naar Maastricht gekomen om dossiers van onze COOP te bestuderen. Het was een hele klus en ook het regelen van ''details'' als een tijdelijke aanstelling was niet gemakkelijk. Maar ik ben heel blij dat jullie ons geholpen hebben bij dit project en bij het schrijven van het bijbehorende artikel. Daarom wil ik bij deze heel graag Rijk Gans, Roelien Enting en Tjalling Waterbolk bedanken.

Stella & Xavier – lieve kamergenoten; bedankt voor de gezelligheid, de fijne gesprekken en het versieren van mijn kamer, ik heb me nog nooit zo jarig gevoeld op mijn verjaardag tijdens het werk als met jullie! :-)

Andere junior KEMTA collega's (Martijn, Svenja, Marije, Willem, Maarten, Ben): sinds de junior chat in het leven geroepen is op whatsapp zijn er al heel wat leuke foto's voorbij gekomen, maar zijn er ook gezellige borrels en etentjes geweest, heel fijn voor de ontspanning!
De rest van de afdeling KEMTA zou ik graag willen bedanken voor het feit dat ieders deur altijd openstaat, voor een onderzoeksvraag maar ook gewoon voor een praatje!

Lieve Eline: ondanks dat we geen collega's meer zijn, zien we elkaar nog regelmatig tijdens onze lunchwandelingen. Maar ook buiten het werk om kijken we de mooiste tranentrekkers in de bioscoop en eten we gezellig samen. Bedankt!

Lieve Outi: als promovendus en daarnaast mama van twee lieve meisjes heb je een druk leven. Door jou besef ik dat alles kan, als je maar wilt! Kiitos!

Lieve Stella: omdat je naast mijn kamergenoot ook mijn paranimf bent, kom je zelfs twee keer voor in dit dankwoord. Dankjewel voor je oog voor detail, hopelijk valt je niks meer op nu het boekje gedrukt is :-). Heel fijn dat je op deze bijzondere dag naast me staat!

Lieve Liv: we zien elkaar niet vaak, maar als we dat wel doen dan is het vanouds gezellig! Iedere keer weer op een andere plek ontmoeten we elkaar om lekker te genieten van de natuur en daarna samen een hapje te eten. Bedankt voor de gezelligheid!

Lieve Nathalie, inmiddels kennen wij elkaar al weer ruim 10 jaar....maar ik kan me de tijd in de bieb van de uni in Wageningen nog goed herinneren! Wat is er veel gebeurd in die jaren zeg, dat hadden we nooit kunnen bedenken. Ik hoop nog heel vaak ''vieze'' thee bij je te drinken en gezellig bij te kletsen!

Dear Denisse: although the distance literally has grown since the beginning of my PhD project, we still have regularly contact via diverse channels. Thanks for being there, always! I hope we will meet again soon!

Lieve Willeke: wat ben ik blij dat jij een van mijn paranimfen wilt zijn en naast me zult staan op dit hele bijzondere moment in mijn leven. Heel erg bedankt voor het doornemen van het manuscript! Vandaag dus eens een keer niet in badpak! Ik hoop dat we nog heel lang samen blijven trainen, ook al komen we terecht in baan 8 :-). Op naar het volgende uitje met z'n tweetjes!

Lieve Renneke: dankzij jou heb ik Venlo al heel wat keren gezien! Onze uitjes zijn altijd heel gezellig en we kunnen altijd zo fijn bijpraten. Zo ben je van een voormalig collega uitgegroeid tot een dierbare vriendin. Onze Facetime-dates waardeer ik enorm, dankjewel! De kaft, uitnodiging en achtergrond van mijn presentatie zijn echt heel mooi geworden. Ik ben heel blij dat ik je de vrije hand heb gegeven!

Lieve Rob en Margareth: bedankt dat jullie gedurende mijn onderzoek regelmatig gevraagd hebben hoe het er mee gaat. En heel fijn dat jullie bij deze bijzondere dag aanwezig willen zijn!

Lieve Lotte en lieve Niels: 2 jaar terug ben je me voorgegaan, Lotte, dankzij jou heb ik van dichtbij ervaren dat promoveren een hele leuke dag is! Helaas wonen we niet dichterbij elkaar, maar met dank aan whatsapp en facetime blijven we wel van elkaars leven op de hoogte!

Lieve Angelika: sinds de komst van Dex kom je iedere donderdag in alle vroegte een heel eind rijden om op hem te passen, samen met Iris. Met in gedachte dat thuis goed op Dex gelet wordt, kan ik in alle rust werken. Met natuurlijk wel een foto of filmpje om te laten zien dat jullie je vermaken! Naast het feit dat ik heel fijn vind dat je komt oppassen, wil ik je bij deze ook bedanken voor het feit dat je iedere week weer zorgt voor een stapel vers gestreken wasgoed! :-)

Lieve mama & lieve papa: ik wil jullie bedanken voor de onvoorwaardelijke steun die jullie me altijd gegeven hebben. Ik kan eindelijk oprecht zijn, als ik nu zeg dat ik trots ben op wat ik bereikt heb. Maar dat had ik niet gekund zonder jullie. Jullie hebben me gemaakt tot wie ik ben! Bedankt dat jullie altijd in me zijn blijven geloven.
Mama, je eerste reactie toen ik zei dat ik ging promoveren, zal ik nooit vergeten: ''oh dan ga je een boek schrijven!'' Het is weliswaar iets dunner dan de boeken die ik normaal gesproken lees, maar wat nu voor jullie ligt mag toch wel een boek genoemd worden. Eindelijk mag ik nu rechtmatig gebruik maken van mijn titel als doctor!

Lieve Danny, in een tijdsspanne van slechts 7 weken zijn we allebei in Roermond geboren en hebben we de eerste jaren van ons leven bij elkaar in de klas gezeten. Daarna zijn we naar andere scholen gegaan en zijn we elkaar uit het oog verloren. Maar het heeft zo moeten zijn; we zijn elkaar weer tegengekomen en sindsdien zijn we samen! <3
Promoveren was iets wat ik al jaren wilde doen en jij hebt me gesteund in die keuze. Afgelopen jaren waren daarnaast hectisch vanwege de komst van ons lieve mannetje Dex en tijdens mijn verlof nog ''even'' een verbouwing en verhuizing. Maar wat is het mooi geworden allemaal! Gelukkig kunnen we nu van alles genieten. Ontzettend bedankt dat je naast de dagelijkse bezigheden ook nog veel tijd besteedt hebt aan de opmaak van mijn proefschrift. Nu heb je dat etentje wel verdiend ;-)

Als laatste nog een bedankje voor de kleinste persoon in mijn leven: lieve Dex! Door jouw komst weten we dat wonderen bestaan. Wat maak je het leven toch mooi! Door jou heb ik geleerd om te genieten van de kleine dingen in het leven. Ik kijk uit naar wat de toekomst nog voor ons drietjes in petto heeft!

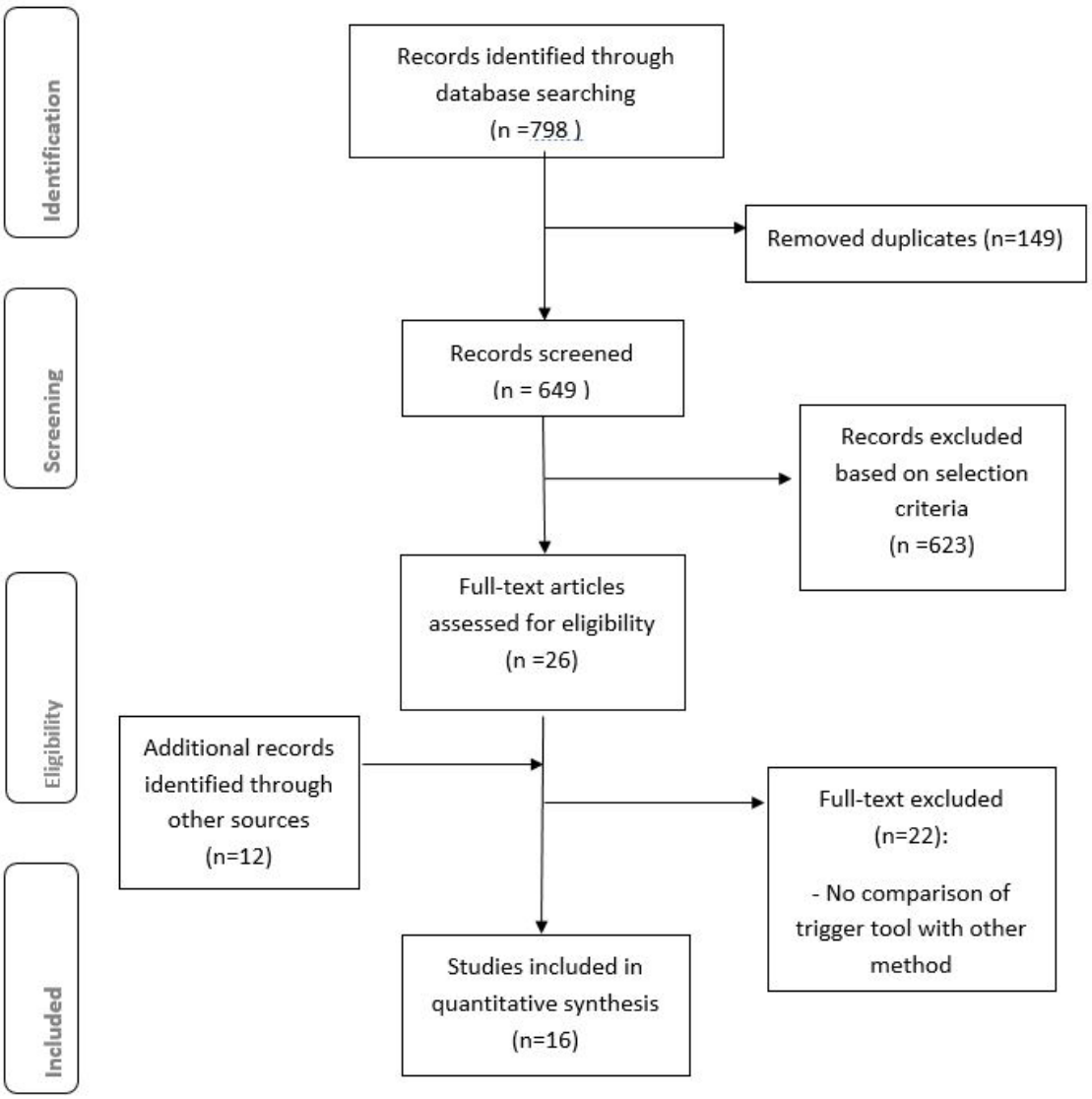
Appendix

Table 1: search strategy

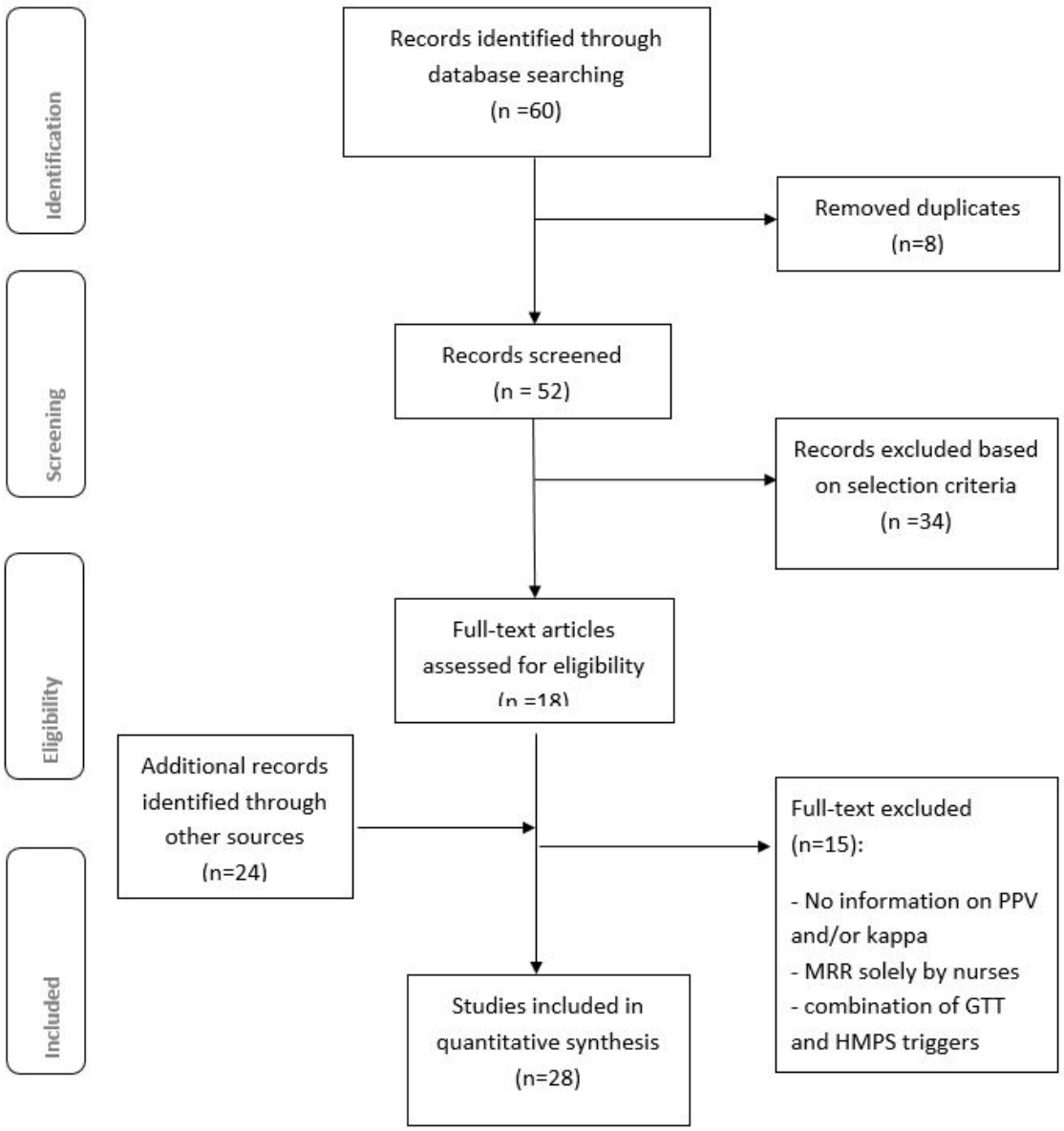
	Database	Search terms
1:	Pubmed Embase Cochrane	(trigger OR GTT OR (global trigger tool) OR (harvard medical practice study) OR HMPS) AND ((adverse event) OR mistake or error) AND (rate or percentage or incidence or prevalence or occurrence or frequency)
2:	Pubmed Embase Cochrane	((inter-rater) OR (intra-rater) OR (inter-observer)) AND ((global trigger tool) or (Harvard medical practice))
3:	Pubmed Embase Cochrane	(Feasible OR valid or validity OR feasibility OR acceptable OR acceptability) AND (medical record review) AND (adverse event)
4:	Pubmed Embase Cochrane	((cost OR costs OR economic* OR financial or finance or (human resources) or burden) AND ((trigger tool) or (global trigger tool) or (Harvard medical practice) or (medical record review)))
5:	Pubmed Embase Cochrane	(medical record review) and ((mission statement) OR ministry OR policy) AND (adverse event)
6:	Pubmed Embase Cochrane	(adverse event) AND ((medical record review) OR (trigger tool) OR (Harvard medical practice)) AND (decreas* OR Improv* OR intervention OR reduction OR reduce OR change*)
7:	Pubmed Embase Cochrane	((medical record review) AND trigger) AND (domain* OR education OR SOP OR guidelines OR (medication safety) OR instrument OR synergy OR outcome* OR effect* or result OR results)

Limit: 2000-2017

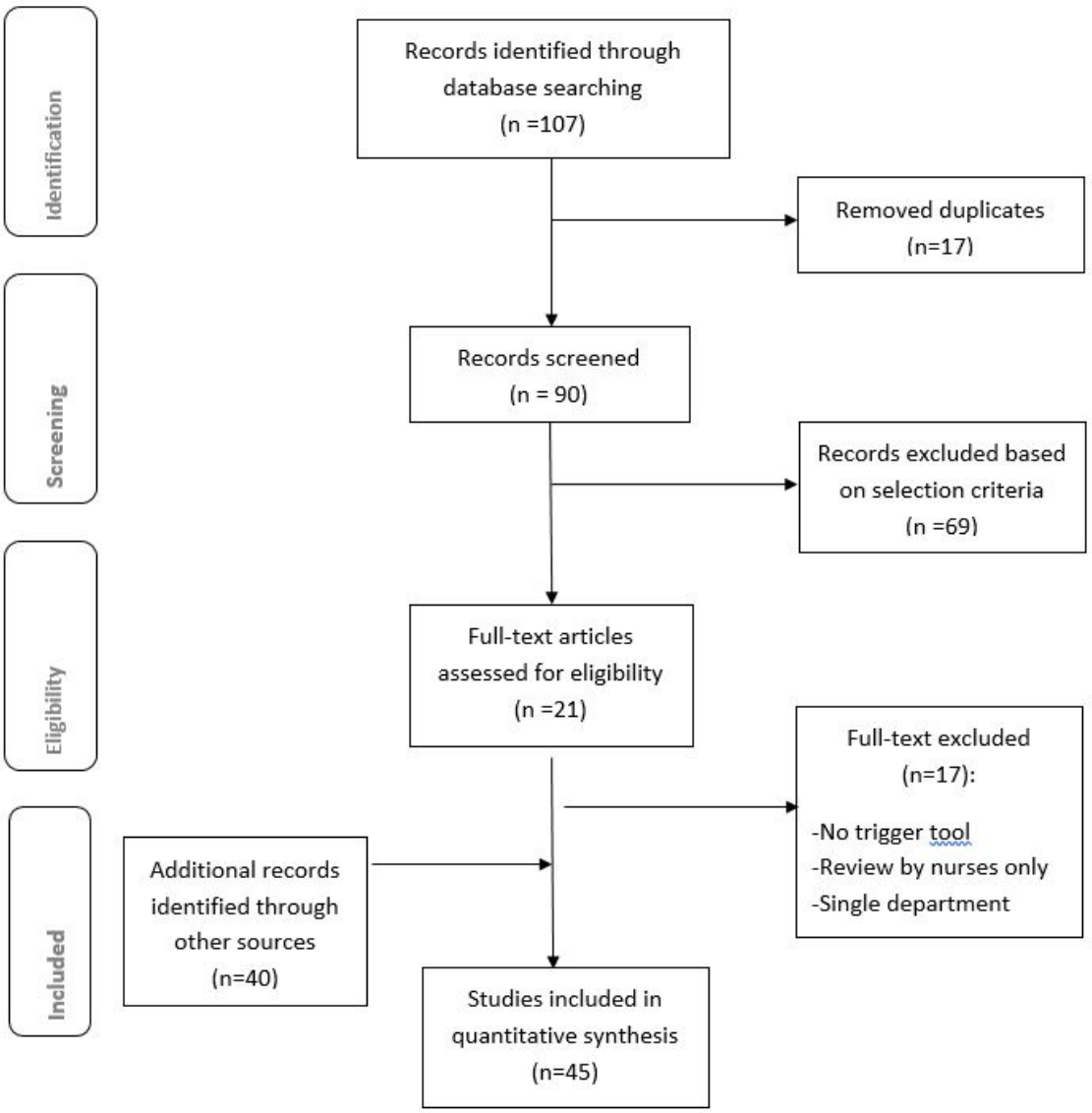
2.1 Flowchart of the literature search and selection of studies criterion 1



2.2 Flowchart of the literature search and selection of studies criterion 2



2.3. Flowchart of the literature search and selection of studies criterion 3



2.4 Flowchart of the literature search and selection of studies criterion 4

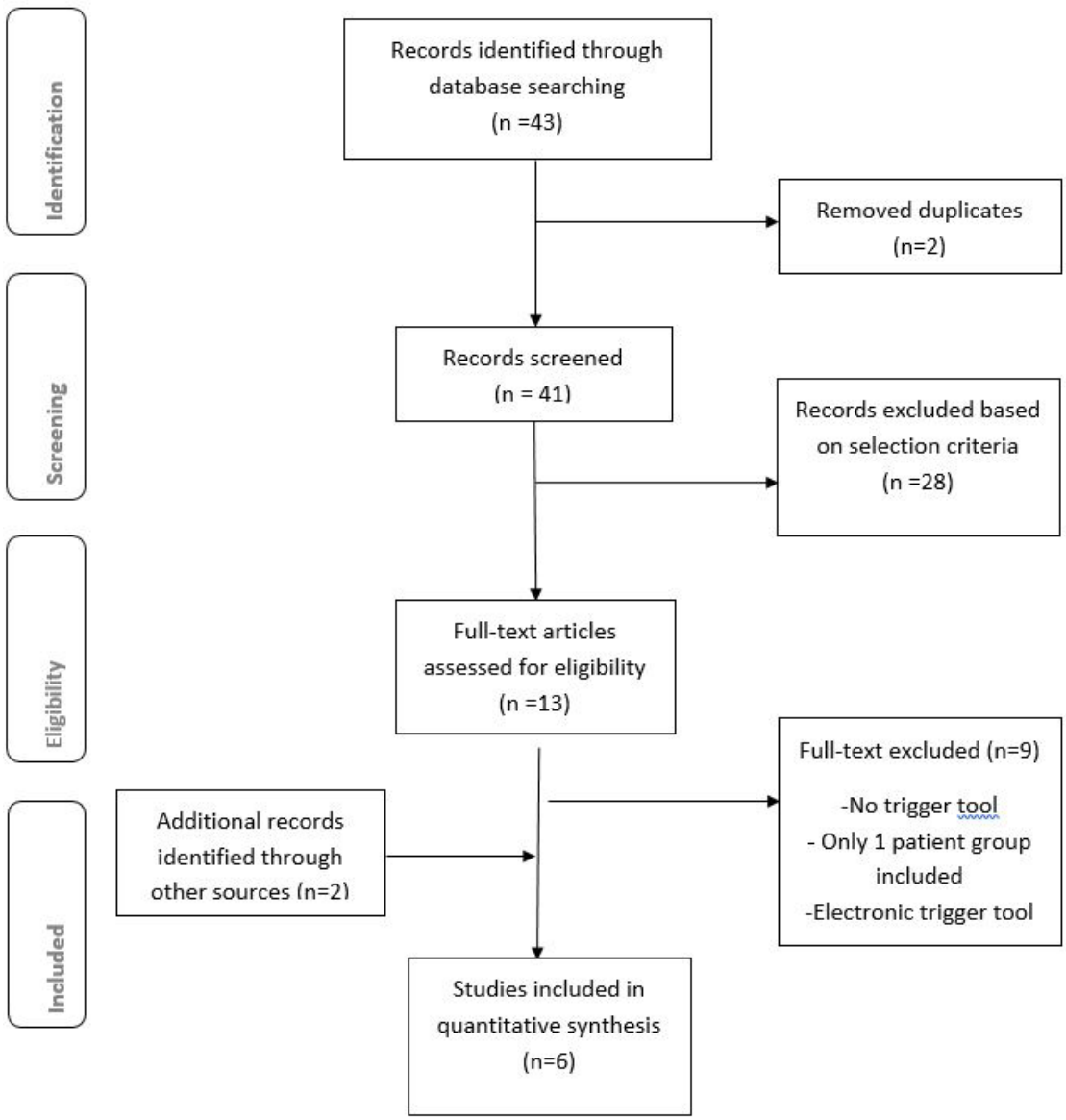


Table 3: Summary of findings

Author	Country	Study Aim	Study design	Instrument	Sample characteristics	Number of screening criteria	Number of reviewers	Number of records to measure reliability	Prevalence AE	Kappa agreement AE (95% CI)	% Agreement	Summary of main results
Adler (2015) [79]	US	The aim of this study was to determine the impact of all-cause inpatient harms on hospital finances and patient clinical outcomes.	retrospective review	GTT	Records of 21,007 inpatients from 24 hospitals between 2009 and 2012.	54	4 analyst 1 physician authenticator	50	12.3	0.69	74	Cost of an AE €1930 (category E). Cost of an AE €4086 (category F-I; 95%CI 4364-). Cost per year: (per hospital) €1,3 million 15,610 (74.3%) experienced no harm, 2818 (13.4%) experienced temporary harm, and 2579 (12.3%) experienced harm.
Asavaroengchai (2009)[76]	TH	Assessing the impact of inpatient harms on hospital finances and patient clinical outcomes.	retrospective review	GTT	576 medical records from five major departments	53	2	576	40.9%	0.86 (0.81-0.90)	NR	236 AEs, 122 preventable (52%). K=0.86 (AEs doctors).
Baines (2015)[103]	NL	Evaluation of three national AE studies	retrospective review	HMPS	15 997 patient admissions were included in the study,	18	1	NA	5.7%	NR	59	The percentage of preventable AEs decreased by 30%, although the overall AE rate didn't change.

Baker (2004) [7]	CA	Description of the frequency and type of AEs in patients admitted to Canadian acute care hospitals and to compare the rate of these AEs across types of hospitals and between medical and surgical care.	retrospective review	HMPS	3745 medical records of 1 teaching, 1 large community and 2 small community hospitals.	18	4 RN 5 MD	10 % of the sample	7.5% (95%CI 5.7-9.3)	0,47 (0.35-0.58)	NR	1527 suspected AEs, 255 confirmed AEs. 94 preventable AEs (3%). PPV=17%
Bjertnaes (2015)[62]	NO	The aim of this study was to test the association between the rates of patient-reported incidents and patient harm documented in the patient record.	retrospective review	GTT and PRIH-I index	10288 hospital admissions from 19 hospitals. Excluding rehabilitation, pediatric and psychiatric health care.	54	47 teams	NR	NR	NR	NR	Patient-reported incidents in hospitals are strongly correlated with patient harm rates based on the GTT. This indicates that patient-reported incidents are related to patient safety, but more research is needed to confirm the usefulness of patient reporting in the evaluation of patient safety.
Blais (2008)[59]	CA	To assess if AES found through MRR are detected by incident reports.	retrospective	incident reports and retrospective chart review HMPS	2213 records of adult inpatients in one of the 19 hospitals	18	6 RN, 3 MD	NA	NR	AE presence: 0.47 (95%CI 0.35-0.58), AE preventability 0.69 (95%CI 0.55-0.83)	91%	Incident reports were present in 15% of the records with AEs and in 4.4% in records without AEs. Sensitivity 15.5 %
Brown (2002)[101]	NZ	Identifying the costs of treating medical injury associated with hospital admissions and the patient characteristics of expensive AEs	cross sec-tional prospective and retrospective review	HMPS	6579 medical records randomly sampled from admissions (in 1998) in 13 general hospitals providing acute care.	18	reported in study Davis	NR	NR	NR	NR	The 850 AEs identified in the NZQHS cost an average of $NZ10.264 per patient. For New Zealand, AEs are estimated to cost the medical system over $NZ870 million, of which over $NZ590 million went toward treating preventable AEs.

Christiaans-Dingelhoff (2011)[94]	NL	Examination to what extent the hospital reporting systems cover the AEs identified by patient record review.	retrospective review & longitudinal	Four reporting systems were linked with the database of reviewed records	5375 medical records of 14 hospitals from 2004. Excluding psychiatry obstetrics and children < 1 yrs.	18	NR number of nurses, 55 MD (2 on each case)	NR	9.3%	NR	NR	Only 18 of the 498 (3.6%) AEs identified by record review were found in one or more of the four reporting systems
Classen (2008)[68]	US	This paper outlines the GTT and evaluates a refined process to improve the interrater reliability of this tool	retrospective review	GTT	Study in which the IHI tool was developed and evaluated.	55	4 nonphysicians, 2 physicians	15	NA	0.397 (0.037-0.757)	73.1%	This study shows that by training the interrater reliability for the number of AEs and their severity improves significantly.
Classen (2011)[57]	US	Evaluation of the ability of three methods among inpatients in 3 hospitals.	retrospective review	GTT, VRNS and internal Voluntary Reporting Notification System (VRNS) and PSIs	795 medical records of patients (also patient deaths)	55	4 RN, 2 MD	NR	44.5	0.847	93.9%	354 AEs found
Davis (2001)[112]	NZ	Assessment the occurrence, impact and preventability of AEs recorded in New Zealand public hospitals.	retrospective review	HMPS	6579 medical records randomly sampled from admissions for 1998 in 13 general hospitals providing acute care.	18		4,116	12,9	0,47	95,5	Evaluation of patient safety in New Zealand.
Davis (2002)[9]	NZ	Assessment of the occurrence and impact of AEs in New Zealand public hospitals	retrospective review	HMPS	6579 medical records randomly sampled from admissions for 1998 in 13 general hospitals providing acute care.	18	1	4,116	12.9	0.47	NR	12.9% (850) of hospital admissions were associated with an AE. Less than 15% was associated with permanent disability or death. AEs added an average of over nine days (median 4 days) to the expected hospital stay.
Deilkås (2015)[104]	NO	Records from 19 hospitals were evaluated.	retrospective review	GTT	Review of 40. 851 medical records	54	19 teams	NR	13.0-16.1%	NR	NR	Estimated AEs rates in severity categories E-I decreased significantly from 16.1% in 2011 to 13.0% in 2013.

Deilkas (2017)[97]	NO, SE	Exploration of similarities and differences in hospital AE rates between Norway and Sweden by reviewing medical records with GTT	retrospective review	GTT	Records were randomly selected from all eligible admissions in 2013.	54	45 teams Norway 63 teams Sweden. 1 or more reviewers. Sweden 1 RN, Norway 2.	NR	13.0 Norway 14.4 Sweden	NR	NR	No significant difference between overall AE rates was found between the two countries. The rate was 13.0% (95% CI 11.7% to 14.3%) in Norway and 14.4% (95% CI 12.6% to 16.3%) in Sweden.
Farup (2015)[51]	NO	This study explored associations between the patient safety culture and AEs, and evaluated the validity of the tools.	retrospective review & longitudinal	Hospital Survey on Patient Safety Culture (HSOPSC) and GTT	272 records (135 department 1 and 137 department 2).	54	one trained team	NR	NR	NR	NR	Dep 1: 10 AEs (7%) Dep 2: 28 (20%). There was an inverse association between the patient safety culture and AEs.
De Feijter (2012)[56]	NL	A more comprehensive overview of medical error in hospitals using different information sources.	retrospective	incident reports, patients complaints and retrospective chart review HMPS	All incident reports for 2007, patients complaints filed in 2007 and retrospective chart review of all inpatients that died in 2008.	18	7 RN, 6 MD	NR	NR	NR	NR	1015 incidents were detected; due to anonymity of patient and staff information, overlap from the different sources couldn't be detected.
Garrett (2013)[84]	US	In this article we describe how AHS, developed and implemented a GTT process, including collection and reporting of the resulting harms data.	retrospective review	GTT	17295 medical records of 25 hospitals were analysed.	NR	NR	NR	85/1000 patient days	NR	78	85 AEs occurred per 1,000 patient days (38 AEs per 100 admissions). A mean of 26% of patients experienced at least one AEs
Good (2011)[83]	US	Adaptation of the IHI to enhance learning from the identified AEs and improve patient safety.	retrospective review	GTT	GTT teams from all hospitals reviewed 40 851 medical records randomly selected from 2 249 957 discharges between 2010-2013	54	external nurse editors	NR	31,1	NR	NR	Based on this sample, AE rates were found to be 68.1 per 1000 patient days, or 50.8 per 100 encounters.

Hoogervorst-Schilp (2015)[102]	NL	To investigate the average and extrapolated excess length of stay and direct costs of AEs (AEs) and preventable AEs in Dutch hospitals	retrospective review	HMPS	2975 patient records of 20 hospitals	18	NR	NR	10,9	NR	NR	Cost of an AE: €2600 (95%CI 1968-3232). Cost of AE Per year: $306 million (national level)
Hwang (2014)[41]	KR	This study aimed to examine the performance of the Global Trigger Tool and to investigate characteristics associated with the occurrence of AEs (AEs).	retrospective review	GTT	random sample of 629 charts.	53	2	60	14.0	0.74	90.0	45 experienced at least one AEs, 61% were preventable. PPV= 16%. K=0,73 (triggers nurses)
Kennerly (2013) [64]	US	To adapt the GTT as a sustainable monitoring tool able to characterize AEs (AEs) for organizational learning, within the context of limited resources.	retrospective review	GTT	Over 4 years, 16,172 medical records were reviewed (8 general acute care hospitals J	53	1-4	94	17.1	0.62	NR	14,184 suspected AEs had positive triggers, 2411 confirmed AEs. PPV= 17%. 12.5% preventable or probably preventable and an additional 59% possibly preventable. K=0,62 (AEs doctors).
Kennerly (2014)[49]	US	To report 5 years of AEs (AEs) identified using an enhanced Global Trigger Tool (GTT) in a large health care system.	retrospective review	AHRQ Patient Safety Indicators (PSIs) and GTT	9,017 Records from monthly random samples of adults	53	external nurse editors	NR	23.6	0.62	NR	2129 AEs were found, 1508 AEs preventable (71%).
Klein (2017)[74]	NL	Evaluation and improvement of the positive predictive value (PPV) of the trigger system in deceased patients	retrospective review	HMPS	The medical records of 2182 patients were investigated	18	NA	NA	26,9	NA	NA	In our sample, the trigger system had an overall PPV for AEs of 47%. More triggers present in a record increased the probability of detecting an AE. Adjustments to the trigger system slightly increased the PPV.

Kobayashi (2008)[93]	JP	Investigate whether AEs recorded in accidents reports can be identified by MRR.	retrospective review	accident reports and HMPS	200 medical records, 100 cases with an accident report and 100 random extra cases	18	1 RN, 1 MD	NA	review A: 25 AEs, review B: 24 AEs	NR	NR	61 of the 100 accident reports had at least one criteria present. Of these 61, 28 had an AE present according to the MRR. Sensitivity of identification of AEs was 89.3% for review A and 85.7% for review B.
Kurutkan (2015)[89]	TR	Measurement of the AE rate and determine the potential usefulness of the GTT	retrospective review	GTT and internal Voluntary Reporting Notification System (VRNS)	208 medical records of a one year period (2012/2013).	53	Eight team members (3 nurses, 2 pharmacists, 1 quality management director, and 2 secretaries)	208	32.2	NR	0.892	GTT is 19 times more sensitive than the VRNS in the AE evaluation process. 67 confirmed AEs, 64% preventable.
Landrigan (2010)[1]	US	Determination of whether state-wide rates of harm have been decreasing over time in North Carolina.	retrospective review	GTT	2341 medical records (collected between 2002 and 2007) excluding pediatric, rehabilitation and psychiatric care.	52	NR	NR (7 tests)	25.1	Internal review teams: 0.64-0.93. external teams: 0.40-0.72	Internal reviewers + experienced reviewers: 81%. External reviewers + experienced reviewers: 75%	588 AEs found, 364 preventable AEs (62%). Inter-rater reliability K=0.64-0.93 internal review K=0.40-0.54 external review Preventability K=0.83 internal review K=0.54 external review Presence of harm K=0.49 Internal vs. experienced K=0.32 External vs. experienced Severity K=0.53 Internal vs. experienced K=0.26 External vs. experienced

Langelaan (2017)[108]	NL	Evaluation of the fourth national AE study in NL	Retrospective review	HMPS	2800 medical records of deceased patients	18	1	280	9.9	NR	54	The amount of care related harm has decreased between 2011/2012 and 2015/2016 but the potentially preventable harm and the potentially preventable death hasn't further decreased but stayed stable.
Macharia (2016)[61]	KE	Comparison of the magnitude and characteristics of inpatient AEs in a tertiary, not-for-profit healthcare facility in Kenya, using medical records review and incident reporting.	retrospective review	HMPS	2,000 records were randomly selected out of 23,026 hospital admission during one year	18	2 RN, 4 MD	43	2,65	The inter-reviewer agreement for the reviewers was moderate at κ=0.40 (95% CI 0.13 - 0.66). The figure improved marginally to κ=0.45 (95% CI 0.38 - 0.74) after adjusting for prevalence and bias.	NR	Review of medical records is preferable to incident reporting in determining the prevalence of AEs in health facilities with limited inpatient quality improvement experience. Further research is needed to determine whether staff education and a positive culture change through promotion of non-punitive UE reporting or a combination of approaches would improve the comprehensiveness of AE reporting.

Martins (2011) [113]	BR	Evaluation the relationship between hospital deaths and adverse events, adjusted for patient risk factors, in hospitalized patients in Brazil.	retrospective review	HMPS	1103 patient charts from hospitalizations in the year 2003 in 3 teaching hospitals in the state of Rio de Janeiro, Brazil	18	4 RN, 2 MD	NR	7,6	NR	NR	The results showed that AEs are not only prevalent, but are associated with serious harm and even death.
Michel (2004)[35]	FR	Comparison of the effectiveness, reliability, and acceptability of estimating rates of AEs and rates of preventable AEs using three methods: cross sectional (data gathered in one day), prospective (data gathered during hospital stay), and retrospective (review of medical records).	cross sectional, prospective and retrospective	HMPS	778 patients: medical (n = 278), surgical (n = 263), and obstetric (n = 237). 37 wards in seven hospitals	17	2 RN, 3 MD	145	NR	(global agreement 91.7%; K = 0.83, 95% confidence interval 0.67 to 0.99), but agreement about preventability was low (67.8%; K= 0.31, 0.05 to 0.57).	92	The prospective and retrospective methods identified similar numbers of medical and surgical cases (70% and 66% of the total, respectively) but the prospective method identified more preventable cases (64% and 40%, respectively). The cross sectional method showed a large number of false positives and identified none of the most serious AEs.

Mortaro (2017)[90]	IT	Implementation of the GTT in Italy	retrospective review	GTT	1320 records in total were evaluated.	53	2 postgraduate students in public health and a physician of the medical board.	2 x 50 records	27,8	The interrater agreement improved significantly after intervention (κ interrater I = 0.52, κ interrater II = 0.80, P < 0.001).	I 68,3 II 87,3	Despite the improvements in the interrater consistency, overall results did not show any significant trend in AEs over time. Future studies may be directed to apply and adapt the GTT methodology to more specific settings to explore how to improve its sensitivity
Mull (2015)[53]	US	A pilot study of the GTT was conducted to assess the rates, types, and harm of AEs detected and to examine the overlap in AE detection between the GTT and other existing methods.	retrospective review	GTT	273 medical records of patients with age > 18 and LOS > 1 days	46	1 RN, 1 MD	NR	NR	NR	NR	Thirteen of the 109 AEs (12%) were also detected by other measures
Naessens (2009)[52]	US	Determine the degree of congruence between several measures of AEs.	retrospective review & cross-sectional	(1) patient safety indicators (PSIs) using ICD-9 diagnosis codes, (2) provider-reported events (3) GTT	All inpatients discharged in 2005 (n = 60 599).	53	NR	NM	NR	NR	NR	65 AEs found. GTT compared with PRE: PPV 14%, NPV 99%. GTT compared with PSI: PPV 3%, NPV 99%.

Naessens (2010)[81]	US	Determination of the inter-rater reliability of the GTT in a practice setting, and exploring the value of individual triggers.	retrospective review	GTT	1138 non-pediatric inpatients from all units across the hospital	55	1	1138	27.0	AE RN-MD 0,71 (95%CI 0.68-0.74) AE RN 0.51 trigger nurses 0.63 (triggers nurses)	NR	913 suspected AEs, 307 confirmed AEs. PPV=34%. K= 0.63 (triggers nurses) K=0.51 AEs nurses. K=0,71 AEs nurses vs. doctors
Najjar (2013) [2]	PS	Evaluation of patient safety levels in Palestinian hospitals and to provide guidance for policymakers involved in safety improvement efforts.	retrospective review	GTT	Random records of 640 discharged patients in 2009.	56	2 nurses 1 physician 1 quality supervisor	640	14.2	RN-RN 0,58 RN-MD 0,89	NR	91 AEs were found, of which 54 (59%) were preventable. 64 (70.4%) resulted in temporary harm, requiring prolonged hospitalization. K=0.58 (triggers nurses); K=0.89 (AEs doctors).
Rafter (2017) [91]	IE	Assessment of the frequency and nature of AEs in Irish hospitals.	retrospective review	HMPS	1574 (53% women, mean age 54 years) randomly selected adult inpatient admissions from a sample of eight hospitals, stratified by region and size		6 RN, 3 physician reviewers	10% of the sample	12,2	0,59	NR	Similar rates to other countries. In a time of austerity, AEs in adult inpatients were estimated to cost over €194 million. These results provide important baseline data on the AE burden and, alongside web-based chart review, provide an incentive and methodology to monitor future patient safety initiatives.
Rutberg (2014)[54]	SE	Description of the level, preventability and categories of AEs (AEs) identified by medical record review using the (GTT). Comparison voluntary AE reporting with medical record reviewing.	retrospective review	GTT	20 randomly selected medical records were reviewed every month from 2009 to 2012 in an university hospital. except the paediatric and psychiatric departments and the obstetric ward,	53	2 RN, 1-2 MD	NR	28.2	NR	NR	Record reviewing identified AEs to a much larger extent than voluntary AE reporting. Healthcare organisations should consider using a portfolio of tools to gain a comprehensive picture of AEs. Substantial costs could be saved if AEs were prevented.

Sari (2007)[60]	GB	Evaluation of the performance of a routine incident reporting system in identifying patient safety incidents	retrospective review	HMPS	1006 hospital admissions in 8 specialties: surgery	18	5 RN 3 physicians	107 stage 1, 90 stage 2	10.9 (95% CI 9,0-12.8)	RN-RN (K= 0.67) and MD-MD (K= 0.76)	90	448 suspected AEs, 303 confirmed AEs. PPV=68%. Inter-rater reliability nurses (K=0,67). Inter-rater reliability doctors (K=0,76).
Sari (2015)[69]	IR	Estimation of the extent, nature and preventability of AEs in Iranian general hospitals.	retrospective review	HMPS	Randomly selected 1200 hospital records from inpatients discharged from each selected hospital between April and September 2012.		x nurses 2 MD	NR	7,3	NR	NR	11% of the patients experience an AE, with 34% preventable. Of these patients, 3.7% develop an AE before they are admitted to hospital and 7.3% of patients develop an AE during their final admission. Adverse drug reactions and operative AEs are more common.
Schildmeijer (2012)[63]	SE	To evaluate agreement in judgement of AEs between well-trained GTT teams from different hospitals	retrospective review	GTT	Five teams from five hospitals conducted reviewed patient records from a random sample of 50 admissions	53	2 RN 1 MD	50	17	MD-MD K=0.45 (0.26 to 0.63) RN-RN K=0,20	92.1	Number of suspected AEs after screening mean teams: 64 Number of confirmed AEs mean teams: 15. Number of preventable AEs; number of mean teams 8. PPV=23 %. K=0,45 (AEs doctors) K=0,20 (triggers nurses)

Sharek (2011)[40]	US	Assessment of the performance characteristics of the Institute for Healthcare Improvement Global Trigger Tool (GTT) to determine its reliability for tracking local and national AE rates.	retrospective review	GTT	10 randomly selected medical records in 10 hospitals in each quarter from January 2002 -December 2007.	54	2-4 RN 2 MD	240 For intra rater reliability a sample of 120 records were reassessed.	24.5% internal review 18.0 % external review	Inter-rater reliability Int: 0.64 Ext: 0.4 Intra-rater reliability Int:0.93 Ext: 0.5 K=0.38 (External) Severity K= 0.26 (Internal) K=0.55 (External) K=0.45 (AEs's doctors)	NR	In this study, the internal teams performed better than the external teams, suggesting that internal teams could form the basis of future large-scale AE studies. Internal review teams are also advantageous with respect to cost, availability of personnel, and the ability to continue local AE measurements over time. A future publication will describe the burden and types of overall AEs (preventable and total), as well the change in rates of AEs over time, in North Carolina.
Soop (2009)[71]	SE	To estimate the incidence, nature and consequences of AEs and preventable AEs in Swedish hospitals.	retrospective review	HMPS	1974 medical records from 28 hospitals. Admissions to psychiatric clinics, rehabilitation, palliative care and day-only admissions were not included.	18	18 RN 17 MD	648	12.3	Before discussed presence of an AE in 91% of the cases (k ¼ 0.80) and upon preventability in 91% of the cases (k= 0.76)..	91	648 suspected AEs, 241 confirmed AEs. 169 preventable AEs (68%). PPV=37%. Inter rater reliability nurses (K=0,53). K=0,8 (AEs doctors), K=0,76 (preventable AEs doctors)

Sousa et al (2014)[73]	PT	Estimation of the incidence of AEs in Portugal	retrospective review	HMPS	1669 medical records	18	6 RN, 5 MD	167	11.1	0,78	NR	Incidence rate of 11.1% AEs, 53.2% were considered preventable. The majority of AEs were associated with surgical procedures (27%), drug errors (18.3%) and hospital acquired infections (12.2%). Most AEs (61%) resulted in minimal or no physical impairment or disability, and 10.8% were associated with death. In 58.6% of the AEs' cases, the length of stay was prolonged on average 10.7 days. Additional direct costs amounted to €470,380.00.
Thomas (2000)[11]	US	Methods similar to the HMPS were used to estimate the incidence and types of AEs and negligent AEs in Utah and Colorado in 1992.	retrospective review	HMPS	15000 patients from all hospitals in Utah and Colorado, nonpsychiatric discharges from 1992.	18	x RN 22 MD	NR	4.0	0,4 (0,3-0,5)	79	587 AEs found, inter-rater reliability of AEs found by doctors K=0.4
Thomas (2002)[114]	US	To measure the reliability of MMR for detecting AEs and negligent AEs.	retrospective review	HMPS	500 medical records out of the sample which contained 15000 records	18	3 MD	500	NR	K=0,40-0,41 for AEs K=0,19-0,24) negligent AEs	NR	Estimates of AE rates are highly dependent on the degree of consensus between the reviewers.
Vincent (2001)[8]	UK	To examine the feasibility of detecting AEs through record review in British hospitals and to make preliminary estimates of the incidence and costs of AEs.	retrospective review	HMPS	1014 patients from two acute care hospitals.	18	4 RN 5 MD	NR	10.8	NR	NR	405 suspected AEs after screening, 119 confirmed AEs. 57 preventable AEs (48%). PPV=29%. Cost of an AE: $3532 additional direct cost for bed days, longer stay on average 8.5 days.

Von Plessen (2012)[92]	DK	To describe experiences with the implementation of global trigger tool (GTT) reviews in five Danish hospitals and to suggest ways to improve the performance of GTT review teams.	retrospective review	GTT	Review teams from five Danish pilot hospitals analysed a sample of the admissions during 18 months from January 2010 until June 2011	56	2 RN 2 MD	NR	NR	NR	NR	We found substantial variation in harm rates. Differences in training, review procedures and documentation in patient records probably contributed to these variations. Training reviewers as teams, specifying the roles of the different reviewers, training records and a database for findings of reviews may improve the application of the GTT.
Weissman (2008)[55]	US	To compare patient reported AE with MRR	retrospective review & interview	HMPS + questionnaire	998 patients were interviewed and their medical records were reviewed for AE	18	2 RN 2 MD	998	12,8	K= 0,85 presence AE, preventability K= 0.71	NR	Although there was some overlap, many AE were present according to the patients, which were not detected by MRR.
Wilson (2012)[75]	EG, JO, KE, MA, ZA, SD, TN, YE	To assess the frequency and nature of AEs to patients in selected hospitals in developing or transitional economies.	retrospective review	HMPS	15548 medical records of 26 hospitals. Same day admissions were not included.	18	1 MD	NR	2,5-18,4	NR	NR	3351 suspected AEs, 1277 confirmed AEs. 485 preventable AEs (38%). PPV=38%. K=0,76 trigger nurses.

Zegers (2009)[5]	NL	This study determined the incidence, type, nature, preventability and impact of AEs (AEs) among hospitalised patients and potentially preventable deaths in Dutch hospitals.	retrospective review	HMPS	7926 admissions: 3983 admissions of deceased hospital patients and 3943 admissions of discharged patients in 2004, of 21 hospitals (4 university, 6 tertiary teaching and 11 general hospitals).	18	66 RN 55 MD	415 phase 1 119 phase 2	5.7% (95% CI 5.1% to 6.4%)	The reliability of the assessment of screening criteria by nurses was good (K=0.62; The reliability of determination of AEs was only fair (k 0.25)	76	4357 suspected AEs, 663 confirmed AEs. 387 preventable AEs (58%). PPV=15%. K=0,64 within pairs.
Zegers (2010)[72]	NL	To evaluate the inter-rater agreement of the record review process of the Dutch AE study, which we aimed to improve by the involvement of two independent physician reviewers per record instead of one including a consensus procedure in case of disagreement.	retrospective review	HMPS	3983 admissions of deceased hospital patients and 3943 admissions of discharged patients in 2004, of 21 hospitals (4 university, 6 tertiary teaching and 11 general hospitals).	18	55 MD	4227	NR	K=0,72 within pairs preventable AEs. K=0,25 between pairs for presence of AE. K=0,40 between pairs preventable Aes	within pairs 91.3% between pairs 75,6 for Aes. For preventable AEs: 86,1 - 70.4	A record review process with two physicians per record including a consensus procedure to assess AEs is not more reliable than a record review process with one physician. Retrospective estimates of incidence of AEs from record review studies should be interpreted with caution. Improvement of the method is necessary for monitoring incidence of AEs over time at a national level.

Table 4: Quality assessment of the included studies

Author	Training	Case selection	Definition of variables	Abstraction forms	Monitoring	Blinding	Interrater agreement discussed	Testing of interrater agreement	Medical record identified	Sampling method	Missing data management	Review board approval	Funding	Sample size calculation	Clear hypothesis and aim
Adler (2015)	1	1	1	0	0	0	0	1	0	1	0	0	0	0	0
Asavaroengchai (2009)	0	1	1	0	0	0	1	1	0	1	0	1	1	0	1
Baines (2015)	1	1	1	1	1	0	0	1	0	1	0	1	1	0	1
Baker (2004)	1	1	1	1	0	0	1	1	0	1	0	1	1	1	1
Bjertnaes (2015)	0	1	1	0	0	0	0	0	0	1	0	1	1	1	1
Blais (2008)	1	1	1	1	0	0	1	1	0	1	0	1	1	0	1
Brown (2002)	1	1	1	0	0	0	0	0	0	1	0	0	0	0	1
Christiaans-Dingelhoff (2011)	1	1	1	1	1	0	0	0	0	1	0	1	1	1	1
Classen (2008)	1	0	1	0	0	0	1	1	0	0	0	0	0	0	1
Classen et al (2011)	1	1	1	0	0	0	0	0	0	1	0	1	1	0	1
Davis et al (2002)	1	1	1	1	0	0	1	1	0	1	0	0	1	0	1
Davis et al (2001)	1	1	1	1	0	0	1	1	0	1	0	0	1	0	1
Deilkås (2017)	1	1	1	0	0	0	0	0	0	1	0	1	1	0	1
Deilkas (2015)	1	1	1	1	0	0	0	0	0	1	0	0	1	0	1
Farup et al (2015)	1	0	1	0	0	0	1	0	0	0	0	1	0	1	1
Feijter 2012	1	1	1	0	0	0	0	0	0	1	0	1	0	0	1
Garrett (2013)	1	1	1	1	0	0	0	1	0	1	0	0	0	0	0
Good (2011)	1	1	0	1	0	0	0	0	0	1	0	0	1	0	0
Hoogervorst-Schilp (2015)	1	1	1	1	1	0	0	0	0	1	0	0	0	0	1
Hwang (2014)	1	1	1	0	0	0	1	1	1	1	0	1	1	1	1
Kennerly (2013)	1	1	1	1	1	0	1	1	0	1	0	0	0	0	1
Kennerly (2014)	1	1	1	1	0	0	0	1	0	1	0	0	1	0	1
Klein (2017)	1	1	1	1	0	0	0	0	0	1	1	1	1	1	1
Kobayashi (2008)	0	1	1	0	0	1	0	0	0	1	0	1	1	0	1
Kurutkan (2015)	1	1	1	0	0	0	1	0	0	1	0	1	1	0	1
Landrigan (2010)	1	1	1	0	1	0	1	1	0	1	0	1	1	0	1
Macharia (2016)	1	1	1	1	0	0	1	1	0	1	0	1	1	1	1
Martins (2011)	0	1	1	0	0	0	0	0	0	1	0	1	0	0	1
Michel (2004)	0	1	1	0	0	0	1	1	0	1	0	0	1	0	1
Mortaro (2017)	1	1	1	1	0	0	1	1	1	1	0	0	0	0	1

Mull (2015)	1	1	1	1	0	0	1	1	0	1	0	1	0	0	1
Naessens (2009)	0	1	1	0	0	0	0	0	0	1	0	1	1	0	1
Naessens (2010)	1	1	1	1	0	0	1	1	0	1	0	1	0	0	1
Najjar (2013)	1	1	1	0	0	0	1	1	0	1	0	1	0	0	1
Rafter (2017)	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1
Rutberg (2015)	0	1	1	0	0	0	0	0	1	1	0	1	1	0	1
Sari (2006)	1	1	0	1	0	0	1	1	0	1	0	1	1	0	1
Sari(2015)	1	1	0	1	0	0	0	0	0	1	0	1	1	0	1
Schildmeijer et al (2012)	1	1	1	1	0	0	1	1	1	1	0	1	1	0	1
Sharek et al (2011)	1	1	1	1	0	0	1	1	0	1	0	1	1	0	1
Soop et al (2009)	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1
Sousa et al (2014)	0	1	1	0	0	0	1	1	0	1	1	1	1	1	1
Thomas et al (2000)	1	1	1	1	1	0	1	1	0	1	0	1	1	0	1
Thomas et al (2002)	1	1	1	1	0	1	1	1	0	1	0	0	1	0	1
Vincent et al (2001)	1	1	1	1	0	0	0	0	0	1	0	0	1	0	1
Von Plessen (2012)	1	1	1	0	0	0	0	0	0	1	0	0	1	0	1
Weisman (2008)	1	1	1	0	0	0	1	1	0	1	0	1	1	0	1
Wilson (2012)	1	1	1	1	0	0	0	1	0	1	1	0	1	0	1
Zegers (2009)	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1
Zegers (2010)	1	1	1	1	1	0	1	1	1	1	0	0	1	0	1

Table 5: PRISMA checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	3 + 4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	4
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	NA
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	4 + 5
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	4
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix S1
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	5
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	5
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	NA
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	NA
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2 for each meta-analysis)	NA